DOCUMENT RESUME

ED 226 007 TM 820 411

AUTHOR Cook, Linda L.; Douglass, James B.
TITLE Analysis of Fit and Vertical Equating with the Three-Parameter Model.
PUB DATE Mar 82
NOTE 77p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982). Some tables are marginally legible due to small print.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS College Entrance Examinations; *Equated Scores; *Goodness of Fit; *Item Analysis; *Latent Trait Theory; *Mathematical Models; Secondary Education; Test Construction; Test Items
IDENTIFIERS National Merit Scholarship Qualifying Test; Preliminary Scholastic Aptitude Test; Scholastic Aptitude Test; *Three Parameter Model; *Vertical Equating


ABSTRACT

The Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT) was equated to the Scholastic Aptitude Test (SAT) using item response theory (IRT) in a three-parameter logistic model. Estimated true formula score equating was used. Only the mathematical sections of the PSAT/NMSQT and SAT were used because the discrepancy in level of difficulty is greater than in the verbal sections and is most representative of a vertical equating situation. The equating used an internal anchor test design. The results of the IRT equating method were compared to those obtained from the conventional linear and curvilinear equating methods. The use of a goodness of fit statistic leading to a chi-square-like test in conjunction with item ability regression plots is examined. The lack of a criterion for judging the equatings and the subjective nature of the goodness of fit assessment are discussed. (CM)

ED226007

TM 820 411

Analysis of Fit and Vertical Equating

With the Three-Parameter Model[1,2]

Linda L. Cook
Educational Testing Service

James B. Douglass
Opinion Research Corporation



[1]A paper presented at the annual meeting of AERA, New York, 1982.

[2]This study was partially supported by the College Board through the funds provided to the PSAT/NMSQT Testing Program.

Analysis of Fit and Vertical Equating

With the Three-Parameter Model

Linda L. Cook
Educational Testing Service

James B. Douglass
Opinion Research Corporation

## Background and Purpose of the Study

Most large scale testing programs are involved in either of two situations that necessitate a statistical process referred to as equating. In the first situation, multiple editions of a test have been constructed to measure a specific aptitude or ability at a defined level of proficiency. The Scholastic Aptitude Test (SAT) administered by Educational Testing Service for the College Board is an example of this type of test. The various editions of the SAT contain different questions but are carefully constructed to be as similar in difficulty and content as possible. In spite of these efforts, it is usually impossible to construct multiple forms that are of exactly the same difficulty level. If some assessment of relative ability is to be made for students taking different editions of the same test, it is critical that a method of equating, rendering comparable the scores on multiple editions of the test, be established. When these multiple editions are testing content at a very similar difficulty level (such as is the case for the SAT) the process is referred to in the literature and in practice as horizontal equating.

In the second situation, the testing program is interested in establishing a single scale that allows scores to be compared for various levels of an aptitude or ability. Typical examples are the many commercially available test batteries that contain tests developed for several grade levels. Because aggregate scores are often compared across levels (e.g. for program evaluation

purposes) it is critical that the scores obtained on the various levels of the test be equated, i.e., placed on a common underlying scale. This type of equating is usually referred to as vertical equating; its purpose being to place on a single scale, multiple editions of a test which are each designed to measure a different level of the same attribute.

In this paper, the vertical equating situation that will be examined is the equating of the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT) to the SAT. The PSAT/NMSQT, similar to the SAT, is administered by Educational Testing Service for College Board and the National Merit Scholarship Corporation. The equating of the PSAT/NMSQT represents a fairly unique situation in that the use of the scores requires the test be equated both vertically and horizontally.

PSAT/NMSQT scores are used for two purposes: (1) to give examinees an idea of what the SAT is like and what scores they might expect to obtain when they take the SAT; and (2) as a means of screening candidates for qualification as National Merit Scholars. The first purpose requires that scores on the PSAT/NMSQT be as comparable as possible to scores on the SAT (a vertical equating situation). The second purpose requires that scores on alternate editions of the PSAT/NMSQT be as comparable as possible (a horizontal equating situation). Only the vertical equating of the PSAT/NMSQT is of interest for this study. The reader is referred to Cook, Dunbar and Eignor (1981) for a discussion of the horizontal equating situation.

There has been a good deal of discussion in the literature as to: (1) the appropriateness of item response theory (IRT) as a method to equate tests; (2) the feasibility of using IRT equating methods in a vertical equating situation; and (3) the choice of IRT model (one-parameter versus three-parameter).

for test equating purposes. The following is a brief review of some of the relevant research.

Slinde and Linn (1977, 1978, 1979) investigated the problem of vertical equating of two forms designed for populations at different levels of ability. Their results suggested that linear, equipercentile and IRT equating employing the one-parameter logistic model may have limitations for the process of vertical equating. This was especially true when the differences between test difficulty and between ability levels of equating samples were most pronounced. Their studies imply that an IRT approach based on the more complex three-parameter logistic model might provide more useful results for vertical equating situations.

Marco, Petersen, and Stewart (1979) presented perhaps the most comprehensive empirical study of equating techniques yet to appear. For designs like the PSAT/NMSQT design, they found problems with traditional methods similar to those found in the Slinde and Linn studies. In particular, when tests differing in difficulty were given to non-equivalent groups and equated using an anchor test design, traditional procedures appeared to break down. In spite of the presence of possible criterion bias confounding some of their results, the authors suggested that the three-parameter logistic model would yield the most acceptable results under unusual or extreme design constraints. However, Marco et al. found, as did Slinde and Linn, that the degree of dissimilarity between groups and test forms were both relevant. When these factors were moderate, traditional methods, both linear and equipercentile, yielded adequate equatings.

A comparison of the stability of results obtained from traditional and IRT procedures was made by Kolen (1981), who used a cross-validation group to

establish a criterion for the evaluation of seven IRT methods and two traditional methods (linear and equipercentile). Kolen, working with the Iowa Tests of Educational Development, had some difficulty evaluating the results obtained from application of the three-parameter logistic model to equate new Level I tests (vocabulary and quantitative thinking tests administered to 9th and 10th graders) and new Level II tests (tests of the same skills administered to 11th and 12th graders) to old tests of vocabulary and quantitative thinking that consisted of one level, administered to grades 9-12. He found that "Although the three-parameter estimated observed score method tended to produce the most stable cross-validation results at Level I of the tests, the results were of only moderate accuracy at Level II. The three-parameter estimated true score equivalents method tended to produce the most stable cross-validation results at Level II and results of moderate stability at Level I." Kolen noted that these results may be partially explained by difficulty in estimating the psuedo-guessing parameter of the three-parameter model and by problems related to the transformation of estimated true scores below the chance score level; he concluded that there may be some doubt regarding the applicability of the three-parameter model in all equating contexts. A somewhat surprising degree of stability was found to hold for conventional equipercentile procedures, sufficient to lead Kolen to conclude that they may be the most viable in practice for designs which involve equating tests differing only moderately in difficulty.

More recently, Petersen, Cook, and Stocking (1981) investigated the relative accuracy of conventional versus IRT equating methods using scale drift as the criterion for evaluating these methods. The authors used a chain of six verbal and six mathematical forms of the SAT, each form linked to the preceeding and following form by an anchor test. The design allowed them to

evaluate the effect of equating a test to itself via five intervening forms. The equating methods investigated were four traditional methods (three linear and one equipercentile) and a three-parameter logistic IRT method that employed three procedures for linking parameter estimates so that estimated true formula scores could be equated.

Results of the study indicated that all three IRT linking methods were superior to the traditional methods for the verbal portion of the test. For the mathematical portion, only one of the IRT methods produced results as satisfactory as those obtained from two of the three linear methods. Equipercentile equating yielded fairly unsatisfactory results for both the verbal and mathematical sections. It should be noted that the study involved tests similar in level of difficulty which were given to groups of examinees that did not differ greatly in their level of ability; a situation in which one would expect traditional linear methods to work well.

If anything, an in-depth look at previous research comparing various equating procedures leaves the practitioner with little confirming evidence. On the one hand, IRT approaches, especially those using the three-parameter logistic model, appear to provide the most accurate results and hence seem appropriate from an empirical perspective as well as a theoretical one. On the other hand, there is some question regarding their stability, although the comparatively small amount of scale drift associated with most of the IRT calibration designs found by Petersen <u>et al.</u> (1981) is evidence in support of their application to parallel forms of aptitude tests administered to groups that are similar in ability. In addition, it is important to note that the studies reviewed indicate that at present the effects of differential relia-

bility and difficulty of test forms and the effect of the non-equivalence of examinee samples do not appear to be completely understood.

The purpose of the present study was to examine the results of applying the three-parameter logistic model to the vertical equating of the PSAT/NMSQT to the SAT. The results of the IRT equating method were compared to those obtained from the conventional equating methods (linear and curvilinear) typically used to equate the tests. In addition, the goodness of fit of the PSAT/NMSQT and SAT data to the three-parameter logistic model was studied.

## Study Design

### Description of the Tests

Both the PSAT/NMSQT and SAT are composed of multiple choice items. The tests differ in length and difficulty, the PSAT/NMSQT contains 65 verbal and 50 mathematical items whereas the SAT contains 85 verbal and 60 mathematical items.

The PSAT/NMSQT consists of two 50-minute sections. The verbal section contains only 5-choice items; the mathematical section contains a mixture of 4- and 5-choice items. Raw scores obtained on the PSAT/NMSQT are most typically transformed to scaled scores on the College Board 200 to 800 scale via linear equating methods. For score reporting purposes, the final digit of the score is dropped. PSAT/NMSQT raw scores are actually formula scores generated from number-right scores using a correction for guessing. Raw scores are computed by the formula $R - kW$, where $R$ is the number of correct responses, $W$ is the number of incorrect responses and $k = 1/A-1$, $A$ being the number of choices per item.

The SAT and its companion test, the Test of Standard Written English (TSWE), consists of six 30-minute sections: two verbal sections, two mathematical sections, one TSWE section, and one experimental section containing an equating test or pretest. The two verbal sections, one mathematical section, and the TSWE contain 5-choice items; the other mathematical section contains a mixture of 4- and 5-choice items. Raw scores on the SAT are also typically transformed to scaled scores on the College Board 200 to 800 scale by linear equating methods[1]. This scale is retained for score reporting. SAT raw scores are formula scores incorporating the correction for guessing procedure previously described.

It should be noted that only the mathematical sections of the PSAT/NMSQT and the SAT were used in this study. This is because the discrepancy in level of difficulty between the PSAT/NMSQT and SAT mathematical sections is greater than the discrepancy in level of difficulty for the corresponding verbal sections. Thus, the equating of the mathematical sections is the most representative of a vertical equating situation.

## Equating Design

The PSAT/NMSQT is equated to the SAT using what is commonly referred to as an internal anchor test design (Angoff, 1971). This design requires administering one form of the total test to one group of examinees, a second form to a second group of examinees and a common set of items (anchor test) to both groups. The anchor test may be included within the total test (internal anchor) or it may be administered separately (external anchor). The anchor test is

[1]IRT equating methods (curvilinear) were introduced in January, 1982.

constructed to be a miniature but otherwise parallel version of the total test and is used to assess the relative ability of the groups of examinees taking the two forms of the test to be equated.

Standard practice in equating new forms of the PSAT/NMSQT is to equate each new form of the test to two old forms of the SAT through separate sets of common items. One can imagine each of the two new PSAT/NMSQT forms produced annually as being composed of three sets of items: (1) items unique to that form; (2) items in common with one old SAT form; and (3) items in common with a second old SAT form. It is important to note that both new forms (Form 1 and Form 2) of the PSAT/NMSQT share items in common with the same two old SAT forms. However, there exists no item overlap between the two new forms, i.e., each new form is equated back to the same two old SAT forms but through different sets of common items. Final scaled scores are determined for each of the PSAT/NMSQT new forms by combining the results obtained from the equatings to the two SAT old forms. In order to permit an examination of the individual equatings, results were not combined for this study.

Figure 1 contains a schematic diagram of the design used to equate the PSAT/NMSQT to the SAT. As mentioned previously, PSAT/NMSQT Form 1 and Form 2 are alternate forms of the PSAT/NMSQT, each containing a subset of items in common with each of the SAT old forms (hereafter designated SAT First Old Form and SAT Second Old Form). Equating samples for all methods contained approximately 3,000 cases selected randomly from data obtained at the regular administration of each of the old and new forms shown in Figure 1. A total of four random samples, one for each of the PSAT/NMSQT new forms and one for each of the SAT old forms, were selected. Table 1 presents sample raw score summary

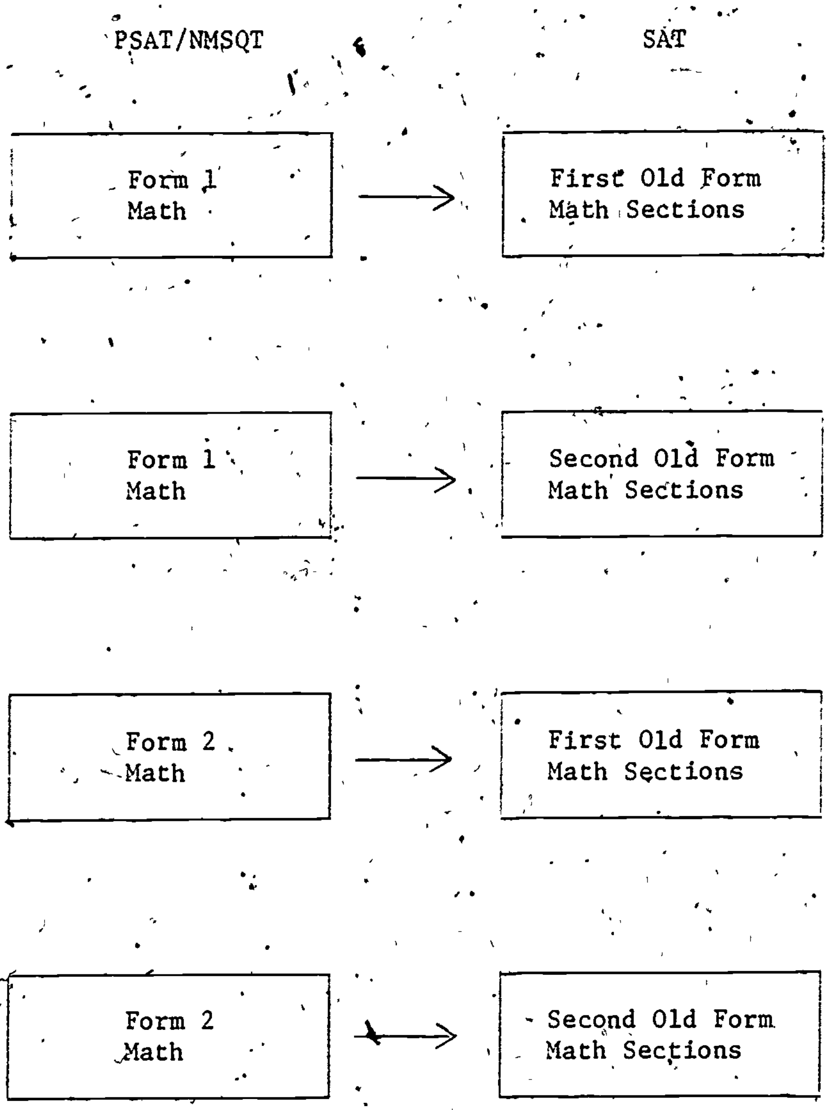



Figure 1: Schematic Diagram of Design Used in Study for Equating PSAT/NMSQT Form 1 and Form 2 Math to SAT First and Second Old Forms.

Table 1

Raw Score Summary Statistics for Equating Samples[1]

| | Form | N | Total Test Mean | Total Test S.D. | Anchor Test Mean | Anchor Test S.D. | Total Test/Anchor Test Covariance |
|---|---|---|---|---|---|---|---|
| First Equating | PSAT/NMSQT Form 1 | 3367 | 19.82 | 11.27 | 7.79 | 4.92 | 51.58 |
| | SAT First Old Form | 3188 | 20.99 | 12.93 | 8.12 | 5.32 | 64.22 |
| Second Equating | PSAT/NMSQT Form 1 | 3367 | 19.82 | 11.27 | 8.30 | 4.84 | 50.57 |
| | SAT Second Old Form | 2763 | 24.47 | 13.18 | 9.36 | 5.06 | 61.34 |
| Third Equating | PSAT/NMSQT Form 2 | 3078 | 22.62 | 11.71 | 8.76 | 5.05 | 55.34 |
| | SAT First Old Form | 3188 | 20.99 | 12.93 | 8.26 | 5.01 | 60.04 |
| Fourth Equating | PSAT/NMSQT Form 2 | 3078 | 22.62 | 11.71 | 8.93 | 4.74 | 51.60 |
| | SAT Second Old Form | 2763 | 24.47 | 13.18 | 9.31 | 4.87 | 59.32 |

[1]The PSAT/NMSQT contains 50 items; the SAT contains 60 items; all anchor tests contain 20 items with the exception of the anchor test used for the fourth equating which contains 19 items.

statistics for the total tests and anchor tests for the four equatings described in Figure 1.

The linear equating methods used in this study both produce an equating transformation of the form $T(x) = Ax + B$, where T is the equating transformation; x is the test score to which it is applied, and A and B are parameters estimated from the data. The Tucker and Levine Unequally Reliable Linear equating models (Angoff, 1971, pp. 579-583) were used in this study.[2] These models are based on univariate selection sampling theory. Scores on the relevant selection attribute (the attribute on which the equating samples vary) are assumed to be collinear with scores on the anchor test in the case of the Tucker model and with true scores on both the anchor test and the test form in the case of the Levine model. Scores on the anchor test are used to estimate performance of the combined group of examinees on both the old and new forms of the test; thus, simulating by statistical methods, the situation in which the same group of examinees take both forms of the test.

The parameters A and B of the equating transformation are estimated by means of an equation that expresses the idea of equating in standard score terms:

$$(x - M_x)/S_x \doteq (y - M_y)/S_y \quad , \qquad (1)$$

where x and y refer to the test scores to be equated, and M and S refer to the means and standard deviations of the scores on some group of examinees. Methods using the above equation differ in their identification of the means and standard deviations to be estimated. The Tucker method is based on the estimated means and standard deviations of observed scores for the combined group whereas the Levine Unequally Reliable method is based on the estimated means and standard

[2] The Levine Unequally Reliable model was used for the PSAT/NMSQT Form 1-SAT Second Old Form equating. All other equatings employed the Tucker linear model.

deviations of true scores for the combined group. The formulas for computing the A and B parameters for the Tucker and Levine Unequally Reliable models are given in Figure 2.

The equipercentile model maintains that scores on two test forms are equivalent if they correspond to the same percentile rank in some group of examinees. The procedure involves equating scores on each test form to the anchor test separately within each group. Scores on the two forms to be equated are then said to be equivalent if they correspond to the same score on the anchor test.

Finally, IRT equating models characterize equivalent scores on two test forms as those scores which correspond to the same estimated level of the latent trait, ability, or skill underlying both tests. Item response theory assumes that a mathematical function relates the probability of a correct response on an item to an examinee's ability (Lord, 1980). The mathematical function (IRT model) employed in this study was the three-parameter logistic model. The model states that the probability of a correct response to item i ($P_i(\theta)$), is given by:

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{1.7a_i(\theta-b_i)}}{1+e^{1.7a_i(\theta-b_i)}}, \quad (i=1,2,\ldots,n) ; \qquad (2)$$

where, $a_i$, $b_i$, and $c_i$ are three parameters describing the item and $\theta$ represents the ability level of an examinee.

The item parameters and examinee abilities for the study were estimated using the program LOGIST (Wood and Lord, 1976; Wood *et al.*, 1976). The estimates are obtained by a (modified) maximum likelihood procedure which has been adapted to accommodate omitted items (Lord, 1974). The following constraints were

Tucker

$$A = (S^2_{yb} + C^2_{yvb}(S^2_{vc} - S^2_{vb})/S^4_{vb})^{\frac{1}{2}}(S^2_{xa} + C^2_{xva}(S^2_{vc} - S^2_{va})/S^4_{va})^{-\frac{1}{2}}$$

$$B = M_{yb} + C_{yvb}(M_{vc} - M_{vb})/S^2_{vb} - AM_{xa} - AC_{xva}(M_{vc} - M_{va})/S^2_{va}$$

Levine Unequally Reliable

$$A = ((S^2_{yb} - S^2_{y''b})/(S^2_{vb} - S^2_{v''b}))^{\frac{1}{2}}((S^2_{xa} - S^2_{x''a})/(S^2_{va} - S^2_{v''a}))^{-\frac{1}{2}}$$

$$B = M_{yb} + (M_{va} - M_{vb})((S^2_{yb} - S^2_{y''b})/(S^2_{vb} - S^2_{v''b}))^{\frac{1}{2}} - AM_{xa}$$

Notation:

| | |
|---|---|
| New Test Form | X |
| Old Test Form | Y |
| Either New or Old Test Form | P |
| Anchor Test | V |
| Observed Score | x, y, v |
| Error Score | x'', y'', v'' |
| Group Taking Test X and Test V | a |
| Group Taking Test Y and Test V | b |
| Group Taking Test P and Test V | g |
| Combined Group | c or (a + b) |
| Mean | M |
| Standard Deviation | S |
| Covariance | C |

Figure 2: Formulas for Linear Conversion Parameters

imposed on the estimation process: a's were restricted to values between 0.01 and 1.75 inclusive; $\theta$'s were restricted to a range of -7.0 to 6.0, and c's were restricted to values between 0.0 and .20. Additionally, each examinee was required to have responded to 17 items in order to insure stable $\theta$ estimates. The constraints were chosen on the basis of previous experience with the data and were imposed to speed convergence of the likelihood function and to maximize stability of the item parameter and ability estimates.

Although a variety of equating techniques exist once an IRT model has been chosen, only estimated true formula score equating (Lord, 1980, Chapter 13) was used for this study. Estimated true formula scores $\hat{\xi}$ and $\hat{\eta}$ on two tests measuring the same ability, $\theta$, are related by the equations,

$$\hat{\xi} = \sum_{i=1}^{n} \hat{P}_i(\theta) - \left[\sum_{i=1}^{n} \hat{Q}_i(\theta)\right] /A-1, \text{ and} \tag{3}$$

$$\hat{\eta} = \sum_{j=1}^{n} \hat{P}_j(\theta) - \left[\sum_{j=1}^{n} \hat{Q}_j(\theta)\right] /A-1 , \tag{4}$$

where, A is the number of choices per item, $\hat{P}_i(\theta)$, and $\hat{P}_j(\theta)$, represent the probability of a correct response for items i and j as they appear in the two forms to be equated and $\hat{Q}_i(\theta)$, $\hat{Q}_j(\theta)$ equal $1 - \hat{P}_i(\theta)$ and $1 - \hat{P}_j(\theta)$, respectively. Using expressions 3 and 4, it is possible to find an estimated true formula score $\xi$ corresponding to an estimated true formula score $\eta$ for any given $\theta$.

Expressions 3 and 4 will not provide equated estimated true formula scores for scores on the two test forms of interest that fall below the chance score level. Several ways exist for determining the relationship in this

region. Kolen (1981) used linear interpolation. The method that was used for this study involved estimating the mean and standard deviation of scores below the chance score level for the two forms of interest and using the estimated values to establish a linear relationship.

The means and standard deviations of below chance score level scores were estimated using the following expressions:

$$M_x = \frac{A}{A-1} \sum_{i=1}^{n_x} \hat{c}_i - \frac{n_x}{A-1}\text{, and} \tag{5}$$

$$S_x^2 = \left(\frac{A}{A-1}\right)^2 \left[\sum_{i=1}^{n_x} \hat{c}_i - \sum_{i=1}^{n_x} \hat{c}_i^2\right], \tag{6}$$

where,

$M_x$ = the mean of PSAT/NMSQT scores below chance level,

$S_x^2$ = the variance of PSAT/NMSQT scores below chance level,

A = the number of choices per item, and

$\hat{c}_i$ = the psuedo guessing parameter for item i.

Equations 5 and 6 were repeated to obtain $M_y$ and $S_y^2$, the estimated mean and variance of below chance level scores for the SAT old form of interest. Linear parameters for equating PSAT/NMSQT scores below chance level to SAT scores below chance level were determined as follows:

$$A = \frac{S_y}{S_x} \tag{7}$$

$$B = M_y - A\,M_x \tag{8}$$

The linear parameters (A and B) are used to form the following expression:

$$\text{score (SAT)} = A\,[\text{score (PSAT/NMSQT)}] + B\,. \tag{9}$$

The item calibration plan for the IRT equatings is illustrated in Figure 3. The entire matrix shown in this figure represents a single LOGIST run. Each of the four groups is conceptualized as taking exactly the same test. The test is conceptualized as containing eight components designated by the column headings. This design places all item parameter estimates on the same scale and permits true formula score equating of each PSAT/NMSQT-SAT pairing shown in Figure 3.

The results of the equatings were evaluated simply by comparing the raw-score to scaled-score transformations (tabled data and graphs) obtained by the three equating methods. Unfortunately, no objective criterion is available in this study to judge the adequacy of the equatings. Most probably the assumptions underlying all of the models have been violated to some extent and their robustness in an anchor test situation is not clearly understood. (See, however, Marco, Petersen and Stewart (1979) and Petersen, Marco and Stewart (in press) for a detailed analysis of the robustness of many anchor-test design methods.) In the absence of a true criterion for judging the equatings, an effort was made to examine the goodness of fit of the data to the IRT model. The method used for this examination is described in the following section.

## Assessment of Goodness of Fit

Researchers often attempt to assess the fit of an item response theory model to real data using a chi-square test or other similar approaches (Wright and Panchapakesan, 1969; Wright and Stone, 1979). The problems associated with this approach have been discussed extensively in the literature (Rentz and Rentz, 1978; Divgi, 1981; Rentz and Ridenour, 1978; McKinley and Reckase, 1980). These problems have both theoretical and practical implications. From a theoretical point of view a problem exists in that chi-square tests require expected

| Group | PSAT/NMSQT Form 1 Unique Items n=10 | PSAT/NMSQT Form 1 - SAT First Old Form Common Items n=20 | PSAT/NMSQT Form 1 - SAT Second Old Form Common Items n=20 | PSAT/NMSQT Form 2 Unique Items n=11 | PSAT/NMSQT Form 2 - SAT First Old Form Common Items n=20 | PSAT/NMSQT Form 2 - SAT Second Old Form Common Items n=19 | SAT First Old Form Unique Items n=20 | SAT Second Old Form Unique Items n=21 |
|---|---|---|---|---|---|---|---|---|
| PSAT/NMSQT Form 1 | X | X | X | Not Reached | Not Reached | Not Reached | Not Reached | Not Reached |
| PSAT/NMSQT Form 2 | Not Reached | Not Reached | Not Reached | X | X | X | Not Reached | Not Reached |
| SAT First Old Form | Not Reached | X | Not Reached | Not Reached | X | Not Reached | X | Not Reached |
| SAT Second Old Form | Not Reached | Not Reached | X | Not Reached | Not Reached | X | Not Reached | X |

Figure 3: Calibration Plan for IRT Equating of PSAT/NMSQT Form 1 and Form 2 Math to the two SAT old Forms. The entire matrix represents a single calibration run. Crosses indicate items that examinee groups were actually exposed to. Each PSAT/NMSQT and SAT sample contains approximately 3,000 cases.

values that are available only when the parameters of the model ($\theta_k$, $a_i$, $b_i$ and $c_i$, in the case of the three-parameter model) are known; in actuality, we have only estimates of these parameters. These estimates are likely to behave differently from the known or true parameters in a statistical test. The practical problems are related to the interpretation of the chi-square values and their associated probability levels. These problems will be discussed in subsequent sections of this paper. One alternative to the various chi-square tests is the use of a graphical technique which involves the comparison of the regression of the observed proportion of people getting an item correct on estimated $\theta$ (empirical regression) with the item response function based on the estimated item parameters (estimated regression) (Hambleton, 1980; Stocking, 1980). The resulting plots are referred to as item ability regressions.

The problem with using item ability regression plots to assess goodness of fit is that the process if fairly subjective. The authors found it quite difficult to examine 141 graphs (one for each item represented in Figure 3) and make consistent judgements regarding the goodness of fit of each item. For this reason, it was decided to use a fit statistic leading to a chi-square like test in conjunction with the item ability regression plots. It should be emphasized that the statistic was used only to aid in the interpretation of the plots. No specific meaning was attached to either the size or the probability levels of the values obtained from the application of the statistic. The fit statistic and the item ability regression plots will each be described briefly in the remainder of this section.

## The Fit Statistic

The fit statistic, referred to as $Q_1'$, is based on a statistic, $Q_1$, suggested by Yen (1981). The two statistics are very similar, the basic

difference being the manner in which examinees are grouped into cells based upon their ability estimates. For both statistics, the initial step is to rank order examinees abilities. For $Q_1$, examinees are divided into 10 cells with approximately equal numbers of examinees in each cell.[2] For $Q_1'$, examinees are divided into 17 cells as follows. Examinees are placed into 15 equally spaced intervals for $\theta$ between +3 and −3. Those examinees with $\theta$ greater than +3 are placed into a single cell and examinees with $\theta$ less than −3 are placed in another cell. Should any cell contain less then 5 examinees, it is collapsed with the adjacent cell closest to $\theta = 0$. The only remaining difference between the two statistics is that for $Q_1'$, the observed proportion of examinees in a particular cell is adjusted for examinees omitting the item. Using Yen's notation, the value of the fit statistic for item i is

$$Q_{1i} = \sum_{j=1}^{17} \frac{N_j (O_{ij} - E_{ij})^2}{E_{ij} (1-E_{ij})} , \tag{10}$$

where,

$N_j$ is the number of examinees in cell j, $O_{ij}$ is the observed proportion of examinees in cell j that passes item i (adjusted for omits) and, $E_{ij}$ is the predicted proportion of examinees in cell j that passes item i,

$$E_{ij} = \frac{1}{N_j} \sum_{k \epsilon j}^{N_j} \hat{P}_i (\hat{\theta}_k) , \tag{11}$$

where $\hat{P}_i(\hat{\theta}_k)$ is the item response function (equation, 2) for item i. It should be noted that the summation is over examinees in cell j. The degrees of freedom are the number of independent data points (cells) less the number of item parameters estimated from these data points. The number of estimated item parameters is not three in all cases. In some instances the value of the item discrimination parameter ($a_i$) was set to the upper bound for a values.[3] In other instances the value of the psuedo-guessing parameter ($c_i$) was set to a common value.[4] Fit statistics were determined using $Q_1$ for each of the 141 PSAT/NMSQT and SAT items used in this study.

Item Ability Regression Plots

The item ability regression plots were obtained as follows. The ability scale ($\hat{\theta}$) is subdivided into 15 equally spaced intervals for a range of -3 to +3. For each interval, equation (12) is used to compute $P_{ij}$, the proportion of people in interval j responding correctly to item i (adjusted for omits). That is,

$$P_{ij} = \frac{N_{ij}^{+} + N_{ij}^{o}/A}{N_{ij}}, \text{ where} \qquad (12)$$

$N_j^+$ is the number of examinees in the jth interval responding correctly to item i,

$N_{ij}^o$ is the number of examinees in the jth interval that omitted item i,

A is the number of alternatives per item,

$N_{ij}$ is the number of examinees in interval j that reached item i.

---

[3]Upper and lower bounds were set for all item discrimination parameters to prevent the estimates from becoming unreasonably large or small.

[4]When LOGIST determines it cannot acurately estimate the c parameter for a certain item, due to insufficient information at lower ability levels, it uses an estimate of c obtained by combining all such items; for this study, c for 5-choice and 4-choice items falling into this category was .067 and .123 respectively.

For each item, 15 P's are plotted as squares whose areas are proportional to $N_{ij}$ (these values constitute the empirical item ability regression). Also plotted with each square is a line of length $4\sqrt{PQ/N_{ij}}$, where P and Q are computed from the estimated item response function. The resulting 15 lines are centered on the estimated item response function which also appears on the plot. It should be noted that although the line is a rough estimate of the .95 confidence interval around the item response function, it is not being used as a statistical test for several reasons: (1) the use of 2 as a coefficient instead of 1.96; (2) the use of the inappropriate symmetric normal approximation to the binomial confidence interval around the response function (particularly a problem for extreme values of P); and (3) the use of an interval based on estimated item parameters. Item ability regression plots were obtained for each of the 141 PSAT/NMSQT and SAT items used in this study.

## Results

### Equating.

Tables 2-5 list the raw-score to scaled-score transformations for each PSAT/NMSQT-SAT pairing for each equating method. The tables also include discrepancies for the linear and equipercentile equating results as compared to the results obtained from the IRT equating. The discrepancies were computed by subtracting the scaled scores obtained using the traditional equating methods from those obtained using the IRT method. The data summarized in Tables 2-5 are presented graphically in Figures 4 and 5. Figure 4 contains plots of the

TABLE 2

RAW SCORE TO SCALED SCORE TRANSFORMATIONS AND RESIDUALS
PSAT/NMSQT FORM 1 TO SAT FIRST OLD FORM

| | | ESTIMATED SCALED SCORE | | | | |
|---|---|---|---|---|---|---|
| RAW SCORE | FREQ | IRT | LINEAR | IRT-LINEAR | EQUIP | IRT-EQUIP |
| 50 | 351 | 78.6 | 72.1 | 6.5 | 77.2 | 1.4 |
| 49 | 884 | 76.8 | 71.2 | 5.6 | 75.7 | 1.1 |
| 48 | 1302 | 75.4 | 70.3 | 5.1 | 74.2 | 1.2 |
| 47 | 746 | 73.9 | 69.3 | 4.6 | 73.0 | .9 |
| 46 | 2602 | 72.4 | 68.4 | 4.0 | 71.8 | .6 |
| 45 | 3686 | 70.9 | 67.5 | 3.4 | 70.6 | .3 |
| 44 | 4889 | 69.4 | 66.5 | 2.9 | 69.4 | 0.0 |
| 43 | 2637 | 68.0 | 65.6 | 2.4 | 68.4 | -.4 |
| 42 | 5236 | 66.6 | 64.6 | 2.0 | 67.4 | -.8 |
| 41 | 7374 | 65.3 | 63.7 | 1.6 | 66.1 | -.8 |
| 40 | 8984 | 63.9 | 62.8 | 1.1 | 64.7 | -.8 |
| 39 | 9570 | 62.7 | 61.8 | .9 | 63.2 | -.5 |
| 38 | 6619 | 61.4 | 60.9 | .5 | 61.9 | -.5 |
| 37 | 11723 | 60.2 | 60.0 | .2 | 60.8 | -.6 |
| 36 | 13816 | 59.0 | 59.0 | 0.0 | 59.7 | -.7 |
| 35 | 15480 | 57.9 | 58.1 | -.2 | 58.5 | -.6 |
| 34 | 13366 | 56.7 | 57.2 | -.5 | 57.3 | -.6 |
| 33 | 13900 | 55.6 | 56.2 | -.6 | 56.0 | -.4 |
| 32 | 18360 | 54.6 | 55.3 | -.7 | 54.8 | -.2 |
| 31 | 20617 | 53.5 | 54.4 | -.9 | 53.6 | -.1 |
| 30 | 21755 | 52.5 | 53.4 | -.9 | 52.4 | .1 |
| 29 | 16390 | 51.5 | 52.5 | -1.0 | 51.5 | 0.0 |
| 28 | 21831 | 50.5 | 51.6 | -1.1 | 50.5 | 0.0 |
| 27 | 24950 | 49.6 | 50.6 | -1.0 | 49.5 | .1 |
| 26 | 26092 | 48.6 | 49.7 | -1.1 | 48.4 | .2 |
| 25 | 23933 | 47.7 | 48.8 | -1.1 | 47.4 | .3 |
| 24 | 19739 | 46.8 | 47.8 | -1.0 | 46.5 | .3 |
| 23 | 26652 | 45.9 | 46.9 | -1.0 | 45.6 | .3 |
| 22 | 28002 | 45.0 | 46.0 | -1.0 | 44.6 | .4 |
| 21 | 28040 | 44.1 | 45.0 | -.9 | 43.5 | .6 |
| 20 | 21193 | 43.2 | 44.1 | -.9 | 42.6 | .6 |
| 19 | 23506 | 42.3 | 43.2 | -.9 | 41.8 | .5 |
| 18 | 27632 | 41.4 | 42.2 | -.8 | 41.0 | .4 |
| 17 | 27738 | 40.6 | 41.3 | -.7 | 40.2 | .4 |
| 16 | 25417 | 39.7 | 40.3 | -.6 | 39.4 | .3 |
| 15 | 18302 | 38.9 | 39.4 | -.5 | 38.8 | .1 |
| 14 | 24898 | 38.0 | 38.5 | -.5 | 38.1 | -.1 |
| 13 | 26335 | 37.2 | 37.5 | -.3 | 37.5 | -.3 |
| 12 | 26252 | 36.3 | 36.6 | -.3 | 36.5 | -.2 |
| 11 | 19074 | 35.5 | 35.7 | -.2 | 35.6 | -.1 |
| 10 | 18440 | 34.6 | 34.7 | -.1 | 34.9 | -.3 |
| 9 | 22863 | 33.8 | 33.8 | 0.0 | 34.2 | -.4 |
| 8 | 23123 | 32.9 | 32.9 | 0.0 | 33.4 | -.5 |
| 7 | 20601 | 32.1 | 31.9 | .2 | 32.7 | -.6 |
| 6 | 12586 | 31.2 | 31.0 | .2 | 32.0 | -.8 |
| 5 | 16949 | 30.4 | 30.1 | .3 | 31.2 | -.8 |
| 4 | 17149 | 29.5 | 29.1 | .4 | 30.5 | -1.0 |
| 3 | 15263 | 28.6 | 28.2 | .4 | 29.8 | -1.2 |
| 2 | 10138 | 27.7 | 27.3 | .4 | 29.0 | -1.3 |
| 1 | 7185 | 26.8 | 26.3 | .5 | 28.3 | -1.5 |
| 0 | 8516 | 25.9 | 25.4 | .5 | 27.4 | -1.5 |
| -1 | 6554 | 25.0 | 24.5 | .5 | 26.7 | -1.7 |
| -2 | 4238 | 24.1 | 23.5 | .6 | 26.2 | -2.1 |
| -3 | 1405 | 23.2 | 22.6 | .6 | 25.7 | -2.5 |
| -4 | 1792 | 22.3 | 21.7 | .6 | 25.1 | -2.8 |
| -5 | 1179 | 21.3 | 20.7 | .6 | 24.4 | -3.1 |
| -6 | 570 | 20.1 | 19.8 | .3 | 23.6 | -3.5 |
| -7 | 231 | 19.0 | 18.8 | .2 | 23.0 | -4.0 |
| -8 | 66 | 18.1 | 17.9 | .2 | 22.4 | -4.3 |
| -9 | 70 | 17.1 | 17.0 | .1 | | |
| -10 | 21 | 16.2 | 16.0 | .2 | | |
| -11 | 3 | 15.2 | 15.1 | .1 | | |
| -12 | 0 | 14.3 | 14.2 | .1 | | |
| -13 | 0 | 13.4 | 13.2 | .2 | | |
| -14 | 0 | 12.4 | 12.3 | .1 | | |

TABLE 3

RAW SCORE TO SCALED SCORE TRANSFORMATIONS AND RESIDUALS
PSAT/NMSQT FORM 1 TO SAT SECOND OLD FORM

| | | ESTIMATED SCALED SCORE | | | | |
|---|---|---|---|---|---|---|
| RAW SCORE | FREQ | IRT | LINEAR | IRT-LINEAR | EQUIP | IRT-EQUIP |
| 50 | 351 | 77.[illegible] | 73.2 | 4.3 | 76.4 | 1.1 |
| 49 | 884 | 75.[illegible] | 72.2 | 3.5 | 74.6 | 1.1 |
| 48 | 1302 | 74.2 | 71.3 | 2.9 | 73.1 | 1.1 |
| 47 | 746 | 72.7 | 70.3 | 2.4 | 71.8 | .9 |
| 46 | 2602 | 71.2 | 69.4 | 1.8 | 70.5 | .7 |
| 45 | 3686 | 69.7 | 68.4 | 1.3 | 69.0 | .7 |
| 44 | 4889 | 68.3 | 67.5 | .8 | 67.5 | .8 |
| 43 | 2637 | 66.9 | 66.5 | .4 | 66.2 | .7 |
| 42 | 5236 | 65.6 | 65.5 | .1 | 65.3 | .3 |
| 41 | 7374 | 64.4 | 64.6 | -.2 | 64.3 | .1 |
| 40 | 8984 | 63.3 | 63.6 | -.3 | 63.4 | -.1 |
| 39 | 9570 | 62.1 | 62.7 | -.6 | 62.4 | -.3 |
| 38 | 6619 | 61.1 | 61.7 | -.6 | 61.5 | -.4 |
| 37 | 11723 | 60.0 | 60.8 | -.8 | 60.7 | -.7 |
| 36 | 13816 | 59.0 | 59.8 | -.8 | 59.8 | -.8 |
| 35 | 15480 | 58.0 | 58.8 | -.8 | 58.9 | -.9 |
| 34 | 13366 | 57.0 | 57.9 | -.9 | 57.8 | -.8 |
| 33 | 13900 | 56.0 | 56.9 | -.9 | 56.8 | -.8 |
| 32 | 18360 | 55.0 | 56.0 | -1.0 | 55.8 | -.8 |
| 31 | 20617 | 54.1 | 55.0 | -.9 | 54.9 | -.8 |
| 30 | 21755 | 53.1 | 54.1 | -1.0 | 54.0 | -.9 |
| 29 | 16390 | 52.1 | 53.1 | -1.0 | 53.2 | -1.1 |
| 28 | 21831 | 51.2 | 52.1 | -.9 | 52.3 | -1.1 |
| 27 | 24950 | 50.3 | 51.2 | -.9 | 51.3 | -1.0 |
| 26 | 26092 | 49.3 | 50.2 | -.9 | 50.3 | -1.0 |
| 25 | 23933 | 48.4 | 49.3 | -.9 | 49.3 | -.9 |
| 24 | 19739 | 47.5 | 48.3 | -.8 | 48.3 | -.8 |
| 23 | 26652 | 46.6 | 47.4 | -.8 | 47.3 | -.7 |
| 22 | 28002 | 45.6 | 46.4 | -.8 | 46.3 | -.7 |
| 21 | 28040 | 44.7 | 45.4 | -.7 | 45.2 | -.5 |
| 20 | 21193 | 43.8 | 44.5 | -.7 | 44.4 | -.6 |
| 19 | 23506 | 42.9 | 43.5 | -.6 | 43.5 | -.6 |
| 18 | 27632 | 42.0 | 42.6 | -.6 | 42.5 | -.5 |
| 17 | 27738 | 41.1 | 41.6 | -.5 | 41.3 | -.2 |
| 16 | 25417 | 40.2 | 40.7 | -.5 | 40.5 | -.3 |
| 15 | 18302 | 39.3 | 39.7 | -.4 | 39.6 | -.3 |
| 14 | 24898 | 38.5 | 38.7 | -.2 | 38.7 | -.2 |
| 13 | 26335 | 37.6 | 37.8 | -.2 | 37.8 | -.2 |
| 12 | 26252 | 36.7 | 36.8 | -.1 | 36.9 | -.2 |
| 11 | 19074 | 35.9 | 35.9 | 0.0 | 36.1 | -.2 |
| 10 | 18440 | 35.0 | 34.9 | .1 | 35.2 | -.2 |
| 9 | 22863 | 34.2 | 34.0 | .2 | 34.3 | -.1 |
| 8 | 23123 | 33.3 | 33.0 | .3 | 33.3 | 0.0 |
| 7 | 20601 | 32.5 | 32.0 | .5 | 32.4 | .1 |
| 6 | 12586 | 31.7 | 31.1 | .6 | 31.5 | .2 |
| 5 | 16949 | 30.9 | 30.1 | .8 | 30.7 | .2 |
| 4 | 17149 | 30.1 | 29.2 | .9 | 30.0 | .1 |
| 3 | 15263 | 29.2 | 28.2 | 1.0 | 29.3 | -.1 |
| 2 | 10138 | 28.4 | 27.3 | 1.1 | 28.7 | -.3 |
| 1 | 7185 | 27.6 | 26.3 | 1.3 | 28.1 | -.5 |
| 0 | 8516 | 26.7 | 25.3 | 1.4 | 27.1 | -.4 |
| -1 | 6554 | 25.9 | 24.4 | 1.5 | 26.1 | -.2 |
| -2 | 4238 | 25.0 | 23.4 | 1.6 | 25.5 | -.5 |
| -3 | 1405 | 24.0 | 22.5 | 1.5 | 24.8 | -.8 |
| -4 | 1792 | 23.1 | 21.5 | 1.6 | 23.7 | -.6 |
| -5 | 1179 | 22.0 | 20.6 | 1.4 | 22.8 | -.8 |
| -6 | 570 | 20.8 | 19.6 | 1.2 | 22.3 | -1.5 |
| -7 | 231 | 19.7 | 18.6 | 1.1 | 21.8 | -2.1 |
| -8 | 66 | 18.7 | 17.7 | 1.0 | 21.4 | -2.7 |
| -9 | 70 | 17.7 | 16.7 | 1.0 | | |
| -10 | 21 | 16.8 | 15.8 | 1.0 | | |
| -11 | 3 | 15.8 | 14.8 | 1.0 | | |
| -12 | 0 | 14.9 | 13.8 | 1.1 | | |
| -13 | 0 | 13.9 | 12.9 | 1.0 | | |
| -14 | 0 | 12.9 | 11.9 | 1.0 | | |

TABLE 4

RAW SCORE TO SCALED SCORE TRANSFORMATIONS AND RESIDUALS
PSAT/NMSQT FORM 2 TO SAT FIRST OLD FORM

| RAW SCORE | FREQ | IRT | LINEAR | IRT-LINEAR | EQUIP | IRT-EQUIP |
|---|---|---|---|---|---|---|
| | | ESTIMATED SCALED SCORE | | | | |
| 50 | 725 | 78.6 | 72.1 | 6.5 | 77.0 | 1.6 |
| 49 | 1712 | 77.3 | 71.1 | 6.2 | 75.3 | 2.0 |
| 48 | 1197 | 75.9 | 70.1 | 5.8 | 73.9 | 2.0 |
| 47 | 1636 | 74.3 | 69.2 | 5.1 | 72.5 | 1.8 |
| 46 | 3344 | 72.7 | 68.2 | 4.5 | 71.2 | 1.5 |
| 45 | 4288 | 71.0 | 67.2 | 3.8 | 69.6 | 1.4 |
| 44 | 5098 | 69.4 | 66.3 | 3.1 | 67.7 | 1.7 |
| 43 | 2534 | 67.8 | 65.3 | 2.5 | 66.8 | 1.0 |
| 42 | 5300 | 66.3 | 64.3 | 2.0 | 65.8 | .5 |
| 41 | 6555 | 64.8 | 63.4 | 1.4 | 64.6 | .2 |
| 40 | 7474 | 63.4 | 62.4 | 1.0 | 63.4 | 0.0 |
| 39 | 7066 | 62.0 | 61.4 | .6 | 62.2 | -.2 |
| 38 | 5999 | 60.7 | 60.4 | .3 | 61.0 | -.3 |
| 37 | 8775 | 59.4 | 59.5 | -.1 | 59.8 | -.4 |
| 36 | 9605 | 58.2 | 58.5 | -.3 | 58.7 | -.5 |
| 35 | 10743 | 57.0 | 57.5 | -.5 | 57.6 | -.6 |
| 34 | 8389 | 55.9 | 56.6 | -.7 | 56.4 | -.5 |
| 33 | 9983 | 54.8 | 55.6 | -.8 | 55.3 | -.5 |
| 32 | 11738 | 53.7 | 54.6 | -.9 | 54.1 | -.4 |
| 31 | 12474 | 52.7 | 53.6 | -.9 | 52.9 | -.2 |
| 30 | 12497 | 51.7 | 52.7 | -1.0 | 51.9 | -.2 |
| 29 | 9840 | 50.7 | 51.7 | -1.0 | 51.1 | -.4 |
| 28 | 12997 | 49.7 | 50.7 | -1.0 | 50.2 | -.5 |
| 27 | 14266 | 48.8 | 49.8 | -1.0 | 49.1 | -.3 |
| 26 | 14556 | 47.9 | 48.8 | -.9 | 47.9 | 0.0 |
| 25 | 12515 | 46.9 | 47.8 | -.9 | 47.0 | -.1 |
| 24 | 11752 | 46.0 | 46.9 | -.9 | 46.1 | -.1 |
| 23 | 14493 | 45.2 | 45.9 | -.7 | 45.2 | 0.0 |
| 22 | 15102 | 44.3 | 44.9 | -.6 | 44.3 | 0.0 |
| 21 | 15127 | 43.4 | 43.9 | -.5 | 43.4 | 0.0 |
| 20 | 11257 | 42.5 | 43.0 | -.5 | 42.5 | 0.0 |
| 19 | 13788 | 41.7 | 42.0 | -.3 | 41.6 | .1 |
| 18 | 14666 | 40.8 | 41.0 | -.2 | 40.7 | .1 |
| 17 | 14695 | 39.9 | 40.1 | -.2 | 39.7 | .2 |
| 16 | 12838 | 39.1 | 39.1 | 0.0 | 38.9 | .2 |
| 15 | 10321 | 38.2 | 38.1 | .1 | 38.1 | .1 |
| 14 | 12863 | 37.4 | 37.2 | .2 | 37.4 | 0.0 |
| 13 | 13244 | 36.5 | 36.2 | .3 | 36.5 | 0.0 |
| 12 | 12458 | 35.6 | 35.2 | .4 | 35.5 | .1 |
| 11 | 9108 | 34.8 | 34.2 | .6 | 34.7 | .1 |
| 10 | 9734 | 33.9 | 33.3 | .6 | 33.8 | .1 |
| 9 | 10728 | 33.0 | 32.3 | .7 | 32.9 | .1 |
| 8 | 10371 | 32.1 | 31.3 | .8 | 31.9 | .2 |
| 7 | 8676 | 31.3 | 30.4 | .9 | 31.0 | .3 |
| 6 | 5687 | 30.4 | 29.4 | 1.0 | 30.1 | .3 |
| 5 | 7507 | 29.5 | 28.4 | 1.1 | 29.2 | .3 |
| 4 | 6946 | 28.6 | 27.5 | 1.1 | 28.4 | .2 |
| 3 | 5678 | 27.7 | 26.5 | 1.2 | 27.5 | .2 |
| 2 | 3398 | 26.8 | 25.5 | 1.3 | 26.7 | .1 |
| 1 | 3039 | 26.0 | 24.5 | 1.5 | 26.0 | 0.0 |
| 0 | 3074 | 25.1 | 23.6 | 1.5 | 25.2 | -.1 |
| -1 | 2169 | 24.2 | 22.6 | 1.6 | 24.2 | 0.0 |
| -2 | 1295 | 23.3 | 21.6 | 1.7 | 23.3 | 0.0 |
| -3 | 491 | 22.5 | 20.7 | 1.8 | 22.6 | -.1 |
| -4 | 608 | 21.6 | 19.7 | 1.9 | 21.9 | -.3 |
| -5 | 365 | 20.7 | 18.7 | 2.0 | 20.7 | 0.0 |
| -6 | 170 | 19.8 | 17.7 | 2.1 | 19.0 | .8 |
| -7 | 48 | 18.8 | 16.8 | 2.0 | 18.1 | .7 |
| -8 | 19 | 17.9 | 15.8 | 2.1 | 17.8 | .1 |
| -9 | 23 | 16.9 | 14.8 | 2.1 | | |
| -10 | 3 | 16.0 | 13.9 | 2.1 | | |
| -11 | 3 | 15.1 | 12.9 | 2.2 | | |
| -12 | 0 | 14.1 | 11.9 | 2.2 | | |
| -13 | 1 | 13.2 | 11.0 | 2.2 | | |
| -14 | 0 | 12.3 | 10.0 | 2.3 | | |

TABLE 5

RAW SCORE TO SCALED SCORE TRANSFORMATIONS AND RESIDUALS
PSAT/NMSQT FORM 2 TO SAT SECOND OLD FORM

| RAW SCORE | FREQ | ESTIMATED SCALED SCORE: IRT | LINEAR | IRT-LINEAR | EQUIP | IRT-EQUIP |
|---|---|---|---|---|---|---|
| 50 | 725 | 77.5 | 72.1 | 5.4 | 76.9 | .6 |
| 49 | 1712 | 76.2 | 71.2 | 5.0 | 75.4 | .8 |
| 48 | 1197 | 74.7 | 70.3 | 4.4 | 74.0 | .7 |
| 47 | 1636 | 73.1 | 69.3 | 3.8 | 72.4 | .7 |
| 46 | 3344 | 71.4 | 68.4 | 3.0 | 70.8 | .6 |
| 45 | 4288 | 69.8 | 67.4 | 2.4 | 69.3 | .5 |
| 44 | 5098 | 68.2 | 66.5 | 1.7 | 67.8 | .4 |
| 43 | 2534 | 66.8 | 65.6 | 1.2 | 66.4 | .4 |
| 42 | 5300 | 65.4 | 64.6 | .8 | 65.0 | .4 |
| 41 | 6555 | 64.0 | 63.7 | .3 | 63.6 | .4 |
| 40 | 7474 | 62.8 | 62.8 | 0.0 | 62.5 | .3 |
| 39 | 7066 | 61.6 | 61.8 | -.2 | 61.5 | .1 |
| 38 | 5999 | 60.5 | 60.9 | -.4 | 60.5 | 0.0 |
| 37 | 6775 | 59.3 | 59.9 | -.6 | 59.5 | -.2 |
| 36 | 9605 | 58.3 | 59.0 | -.7 | 58.5 | -.2 |
| 35 | 10743 | 57.2 | 58.1 | -.9 | 57.5 | -.3 |
| 34 | 8389 | 56.2 | 57.1 | -.9 | 56.6 | -.4 |
| 33 | 9983 | 55.2 | 56.2 | -1.0 | 55.7 | -.5 |
| 32 | 11738 | 54.2 | 55.2 | -1.0 | 54.8 | -.6 |
| 31 | 12474 | 53.3 | 54.3 | -1.0 | 53.8 | -.5 |
| 30 | 12497 | 52.3 | 53.4 | -1.1 | 52.8 | -.5 |
| 29 | 9840 | 51.4 | 52.4 | -1.0 | 51.9 | -.5 |
| 28 | 12997 | 50.4 | 51.5 | -1.1 | 50.9 | -.5 |
| 27 | 14266 | 49.5 | 50.5 | -1.0 | 50.0 | -.5 |
| 26 | 14556 | 48.6 | 49.6 | -1.0 | 49.0 | -.4 |
| 25 | 12515 | 47.7 | 48.7 | -1.0 | 48.1 | -.4 |
| 24 | 11752 | 46.7 | 47.7 | -1.0 | 47.3 | -.6 |
| 23 | 14493 | 45.8 | 46.8 | -1.0 | 46.4 | -.6 |
| 22 | 15102 | 44.9 | 45.8 | -.9 | 45.6 | -.7 |
| 21 | 15127 | 44.0 | 44.9 | -.9 | 44.8 | -.8 |
| 20 | 11257 | 43.1 | 44.0 | -.9 | 44.0 | -.9 |
| 19 | 13288 | 42.2 | 43.0 | -.8 | 43.1 | -.9 |
| 18 | 14666 | 41.3 | 42.1 | -.8 | 42.0 | -.7 |
| 17 | 14695 | 40.5 | 41.1 | -.6 | 40.9 | -.4 |
| 16 | 12838 | 39.6 | 40.2 | -.6 | 39.8 | -.2 |
| 15 | 10321 | 38.7 | 39.3 | -.6 | 38.9 | -.2 |
| 14 | 12363 | 37.8 | 38.3 | -.5 | 38.0 | -.2 |
| 13 | 13244 | 36.9 | 37.4 | -.5 | 37.1 | -.2 |
| 12 | 12458 | 36.0 | 36.4 | -.4 | 36.3 | -.3 |
| 11 | 9108 | 35.2 | 35.5 | -.3 | 35.4 | -.2 |
| 10 | 9734 | 34.3 | 34.6 | -.3 | 34.6 | -.3 |
| 9 | 10728 | 33.4 | 33.6 | -.2 | 33.8 | -.4 |
| 8 | 10371 | 32.6 | 32.7 | -.1 | 33.0 | -.4 |
| 7 | 8676 | 31.7 | 31.7 | 0.0 | 32.1 | -.4 |
| 6 | 5687 | 30.9 | 30.8 | .1 | 31.1 | -.2 |
| 5 | 7507 | 30.0 | 29.9 | .1 | 30.3 | -.3 |
| 4 | 6946 | 29.2 | 28.9 | .3 | 29.5 | -.3 |
| 3 | 5678 | 28.4 | 28.0 | .4 | 28.7 | -.3 |
| 2 | 3398 | 27.6 | 27.0 | .6 | 28.0 | -.4 |
| 1 | 3039 | 26.7 | 26.1 | .6 | 27.3 | -.6 |
| 0 | 3074 | 25.9 | 25.2 | .7 | 26.6 | -.7 |
| -1 | 2169 | 25.0 | 24.2 | .8 | 25.8 | -.8 |
| -2 | 1295 | 24.2 | 23.3 | .9 | 25.1 | -.9 |
| -3 | 491 | 23.3 | 22.3 | 1.0 | 24.2 | -.9 |
| -4 | 608 | 22.4 | 21.4 | 1.0 | 23.3 | -.9 |
| -5 | 365 | 21.4 | 20.5 | .9 | 22.8 | -1.4 |
| -6 | 170 | 20.4 | 19.5 | .9 | 22.4 | -2.0 |
| -7 | 48 | 19.5 | 18.6 | .9 | 21.8 | -2.3 |
| -8 | 19 | 18.5 | 17.6 | .9 | 21.2 | -2.7 |
| -9 | 23 | 17.6 | 16.7 | .9 | | |
| -10 | 3 | 16.6 | 15.8 | .8 | | |
| -11 | 3 | 15.7 | 14.8 | .9 | | |
| -12 | 0 | 14.7 | 13.9 | .8 | | |
| -13 | 1 | 13.8 | 13.0 | .8 | | |
| -14 | 0 | 12.8 | 12.0 | .8 | | |

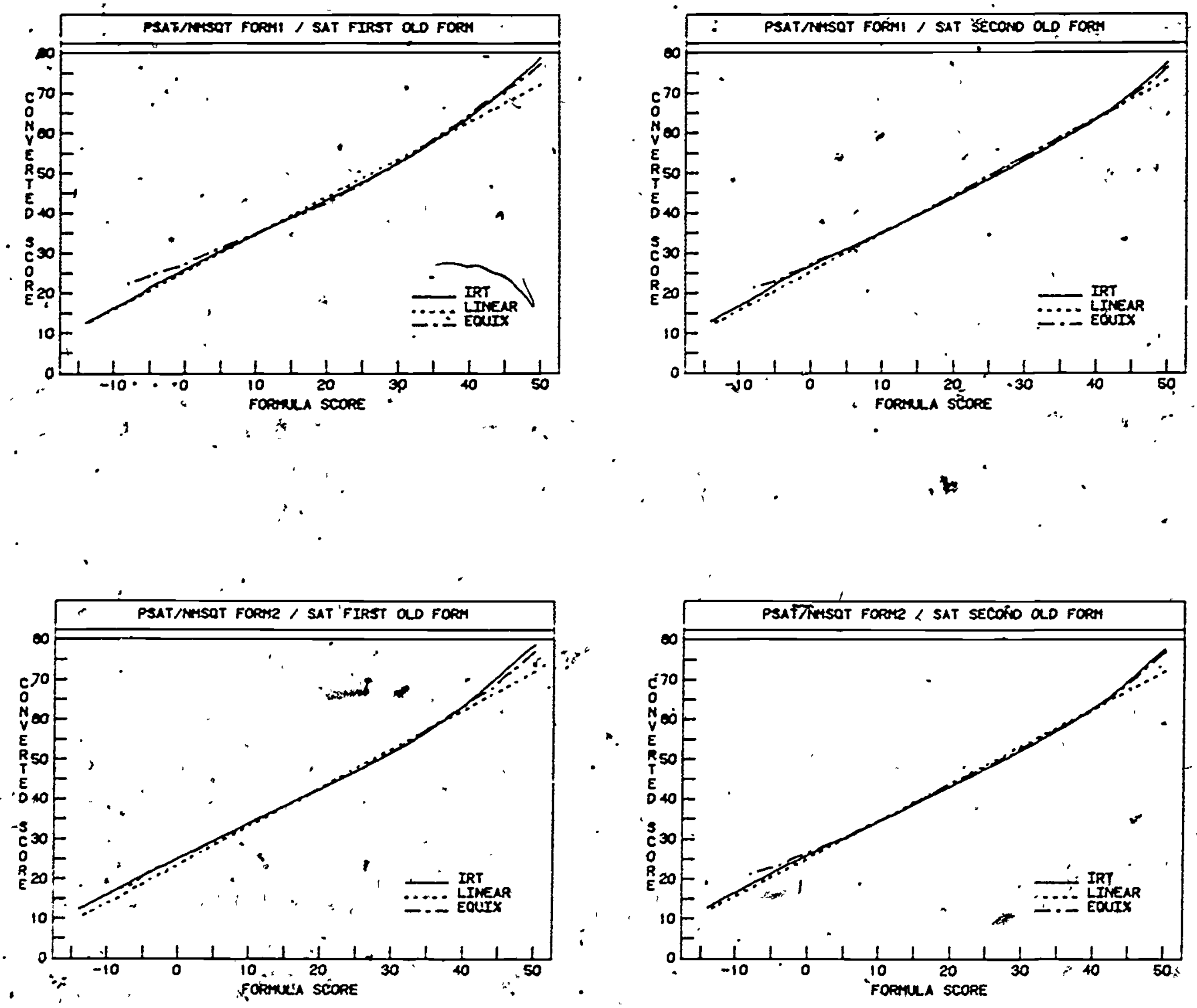



Figure 4: Plots of raw score to scaled score transformations resulting from application of the three equating methods.

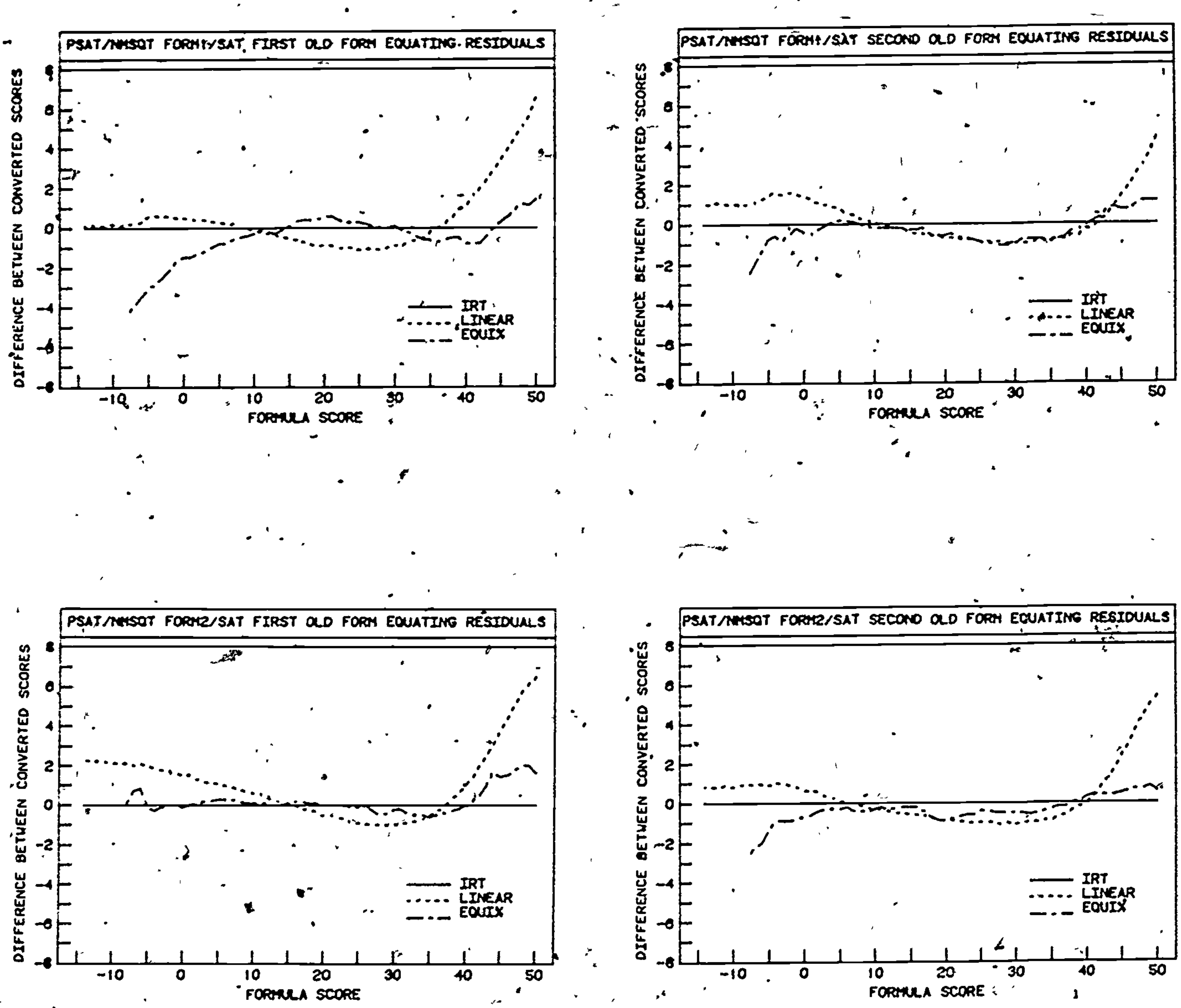



Figure 5: Plots of differences between converted scores (IRT-conventional method) resulting from application of three equating methods.

raw-score to scaled-score transformations obtained for the four equatings. Each plot in this figure compares the results of the conventional equating methods (linear and equipercentile) with the results of the IRT equating.

Plots such as those shown in Figure 4 tend to emphasize the similarities between the equatings rather than the differences. Plots of residuals are often informative when used in conjunction with plots of raw-score to scaled-score transformations. Figure 5 contains residual plots for the four equatings. The IRT equating was used as the comparison equating and the difference between scaled scores obtained by each of the conventional methods and the IRT method was plotted against raw score.

Finally, the effect on scaled score summary statistics of each equating method is presented in Table 6. The data presented in Tables 2-5 were used to compute these statistics. The frequency distributions given in Tables 2-5 are simply a convenient vehicle for converting the scaled scores into interpretable summary statistics. Any reasonable frequency distribution would suffice. The distributions selected were the PSAT/NMSQT Form 1 and Form 2 populations.

Examination of the data found in Table 2 and illustrated in Figures 4 and 5 indicates closer agreement (for scores above a formula score of 10) between the IRT and equipercentile methods than the IRT and linear methods for the PSAT/NMSQT Form 1 - SAT First Old Form pairing. The IRT method tended to yield higher scaled scores at the upper end of the score scale than either of the two conventional methods. The IRT transformations are slightly higher than the linear transformations at the lower end of the score scale and have a tendency to be lower than these transformations through the portion of the score range where the largest number of obtained scores occur. Fairly close agreement between the IRT and equipercentile equating is observed for the middle portion

Table 6

Scaled Score Summary Statistics Resulting from Application of Three Equating Methods

| | Form | N | IRT Mean | IRT S.D. | Equipercentile Mean | Equipercentile S.D. | Linear Mean | Linear S.D. |
|---|---|---|---|---|---|---|---|---|
| First Equating | PSAT/NMSQT Form 1-SAT First Old Form | 828815 | 43.77 | 10.60 | 43.94 | 10.42 | 44.10 | 10.54 |
| Second Equating | PSAT/NMSQT Form 1-SAT Second Old Form | 828815 | 44.21 | 10.38 | 44.66 | 10.53 | 44.50 | 10.79 |
| Third Equating | PSAT/NMSQT Form 2-SAT First Old Form | 462551 | 45.53 | 11.41 | 45.52 | 11.38 | 45.43 | 11.32 |
| Fourth Equating | PSAT/NMSQT Form 2-SAT Second Old Form | 462551 | 45.90 | 11.09 | 46.27 | 10.93 | 46.34 | 10.97 |

of the score range. Noteworthy discrepancies occur at the lower end of the score scale where the equipercentile method yields higher scaled scores than the IRT method.

The differences between the raw-score to scaled-score transformations resulting from the different equating methods applied to the PSAT/NMSQT Form 1 - SAT First Old Form pairing are reflected in the summary statistics found in Table 6. Scaled score means for the IRT and equipercentile methods agree fairly closely. Mean scores resulting from the linear method are slightly higher than those obtained using the IRT equating, reflective of the lower IRT transformations obtained for the major portion of the score reporting range.

Table 3 and Figures 4 and 5 contain information pertaining to the PSAT/NMSQT Form 1 - SAT Second Old Form pairing. Similar to the results of the equating to the SAT First Old Form, the IRT method tends to yield higher scaled scores than either of the conventional methods for the upper end of the score scale. The IRT scaled scores are again lower than the linear scaled scores for the mid-portion of the score range which contains the majority of observed scores. IRT transformations do not agree as closely with the linear transformations obtained for lower scores as they did for the equating to the SAT First Old Form. In this case the IRT transformations are somewhat higher in relationship to the linear transformations as compared to the first situation. There is a slightly greater discrepancy between the IRT and equipercentile transformations for this equating than that found for the equating to the SAT First Old Form. In this instance, the IRT method tended to produce scaled scores that were consistently lower than those produced by the equipercentile method for all scores except those at the upper end of the score range.

Examination of the summary statistics for this equating contained in Table 6 indicates a closer agreement between the scaled score means obtained from the linear and equipercentile methods than the scaled score mean obtained from the IRT method and either the linear or equipercentile method. As was the case with the SAT First Old Form pairing, the IRT scaled score mean is lower than that obtained from either the linear or equipercentile methods.

Information relating to the equating of PSAT/NMSQT Form 2 to the SAT First Old Form is summarized in Table 4 and Figures 4 and 5. Once again, it can be noted that the IRT transformations are higher than either the linear or equipercentile transformations for the upper end of the score scale. It appears that for the mid-portion of the score range the IRT transformations are lower than the linear transformations but not for as great a range as previously observed. IRT transformations become higher than linear transformations at a point where a large number of observed scores are still occurring and continue to be higher than the linear transformations to the bottom of the distribution of observed scores. The IRT transformations agree fairly closely with the equipercentile transformations with the exception of the upper end of the score range.

Reference to the data for this equating contained in Table 6 indicates that for the first time the IRT scaled score mean is slightly higher than that obtained by the linear method. An interesting point to note is the close agreement between the scaled score means obtained from all three of the methods.

Table 5 and Figures 4 and 5 summarize the results of the PSAT/NMSQT Form 2-SAT Second Old Form equating. The typical pattern of the IRT equating resulting in higher scaled scores for the upper end of the distribution is observed, however, this discrepancy is not as great as previously observed for the IRT-equipercentile comparison. The IRT method results in lower scaled

scores for the mid-portion of the score range when compared to scaled scores obtained by the linear method. Equipercentile scaled scores are higher than the IRT scaled scores for all scores except those in the very upper portion of the score range. As indicated by the data for this equating given in Table 6, the IRT scaled score mean is lower than either the linear or equipercentile scaled score mean.

To summarize, in general, IRT equating, when compared to linear or equipercentile methods, yields higher scaled score values for the highest raw scores and to some extent, (particularly when compared to the linear equating results) yields higher scaled score values for the lower raw scores. In the mid-portion of the score range IRT transformations tend to be slightly lower than those obtained by linear or equipercentile methods.

Assessment of Goodness of Fit

The results of the goodness of fit analyses are presented in Table 7 and Figures 6-31. Table 7 contains the overall value of the chi-square statistic as well as the contribution to this statistic of each of the 17 ability level intervals for the 141 items. Inspection of the individual contributions to the overall value of $Q_1'$ for an item is important because in some instances only one interval may contribute close to, or over, half of the total chi-square value. This situation generally occurs for intervals at the extremes of the ability continuum where the number of examinees in the interval is small, thus only a few deviant response patterns may cause the interval to contribute greatly to the total chi-square value. Under these circumstances, simply using the overall value to assess the goodness of fit of the item would be inappropriate.

Table 7 (cont'd)

| Item Parameters | Item Numbers 25 | 26 | 27 | 28 | 29 | 30 | 31 * | 32 | 33 | 34 | 35 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.8915 | 0.9420 | 0.6449 | 0.7727 | 1.3525 | 1.1607 | 0.7028 | 0.5636 | 1.2336 | 0.8148 | 0.8458 | 0.7010 |
| B | 0.6877 | 0.7042 | 0.4812 | 1.0877 | 1.0347 | 1.6635 | -2.0581 | -1.1295 | -0.3188 | -0.5085 | -0.4746 | -0.1348 |
| C | 0.1547 | 0.2230 | 0.0251 | 0.1155 | 0.1055 | 0.1517 | 0.0673 | 0.0673 | 0.1344 | 0.0673 | 0.1631 | 0.1313 |
| Ability Categories > 3. | + 0.18 | + 0.16 | + 0.54 | + 0.60 | + 0.07 | - 0.78 | + 0.03 | - 2.50 | + 0.01 | + 0.09 | + 0.07 | - 2.17 |
| 2.8 | - 0.18 | - 0.26 | - 0.09 | + 0.51 | - 1.16 | + 0.47 | - 12.39 | - 0.45 | + 0.03 | + 0.24 | - 1.99 | - 0.21 |
| 2.4 | - 0.43 | + 1.47 | - 3.05 | - 0.03 | - 0.31 | - 0.85 | + 0.26 | - 3.00 | - 2.32 | - 1.21 | + 0.77 | + 0.04 |
| 2.0 | - 1.12 | - 0.38 | - 0.14 | - 0.19 | + 0.09 | - 1.12 | - 13.90 | + 2.11 | + 0.00 | - 0.71 | - 1.21 | - 0.11 |
| 1.6 | - 0.48 | + 0.05 | + 0.02 | - 0.00 | + 0.09 | + 1.44 | - 0.20 | + 2.81 | - 0.36 | - 0.45 | - 0.09 | + 0.42 |
| 1.2 | + 6.94 | - 0.00 | + 0.26 | - 1.01 | + 0.49 | + 0.01 | - 2.28 | + 3.07 | - 3.52 | + 0.62 | + 2.11 | + 1.22 |
| .8 | + 1.26 | - 0.31 | - 0.63 | + 1.80 | + 0.09 | + 0.17 | - 0.19 | + 0.12 | + 0.02 | + 1.09 | - 0.16 | + 0.71 |
| .4 | - 0.74 | - 0.07 | + 0.93 | + 0.65 | - 3.73 | + 0.00 | - 0.16 | - 0.01 | + 0.03 | - 3.19 | + 1.75 | - 1.50 |
| .0 | - 3.95 | + 0.23 | - 0.10 | - 1.26 | + 0.71 | - 2.34 | + 0.23 | - 0.91 | + 2.55 | + 0.01 | + 0.51 | - 0.77 |
| -.4 | - 0.06 | - 0.05 | + 0.63 | - 0.04 | + 0.00 | - 1.51 | + 0.31 | - 0.22 | - 0.01 | + 1.79 | - 0.73 | - 0.10 |
| -.8 | + 1.78 | - 0.01 | - 0.09 | - 0.34 | + 1.17 | + 1.40 | + 4.53 | - 4.10 | - 2.19 | - 0.09 | + 0.17 | + 0.67 |
| -1.2 | + 2.42 | + 1.30 | - 1.14 | + 0.00 | + 1.75 | + 7.44 | + 1.78 | - 0.00 | - 0.11 | + 0.03 | - 0.66 | - 0.01 |
| -1.6 | + 0.08 | + 1.21 | - 0.00 | + 0.03 | - 1.78 | - 0.07 | - 1.51 | + 3.83 | + 1.54 | - 0.01 | - 0.00 | + 0.45 |
| -2.0 | + 0.08 | + 0.22 | - 0.00 | + 2.99 | + 0.41 | + 0.10 | - 8.81 | + 2.91 | - 0.15 | - 2.41 | + 7.29 | + 0.71 |
| -2.4 | + 0.26 | + 0.28 | + 0.33 | + 4.75 | - 0.04 | + 6.16 | - 0.07 | + 0.30 | + 1.22 | - 0.28 | + 0.02 | + 1.91 |
| -2.8 | - 0.09 | + 5.43 | + 0.13 | + 0.01 | + 3.24 | + 0.01 | + 0.03 | + 0.14 | + 1.09 | + 0.05 | + 0.99 | - 1.96 |
| <-3. | - 0.28 | + 0.14 | - 1.33 | + 0.12 | + 0.48 | - 0.14 | - 1.75 | - 0.13 | + 5.77 | + 0.01 | + 0.29 | + 0.00 |
| TOTAL CHI | 20.33 | 11.57 | 9.43 | 14.32 | 15.60 | 24.00 | 48.44 | 26.62 | 20.92 | 12.27 | 18.83 | 12.96 |

| Item Parameters | Item Numbers 37 | 38 | 39 * | 40 | 41 | 42 | 43 | 44 * | 45 | 46 * | 47 | 48 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.8192 | 0.8368 | 1.0499 | 0.4337 | 1.1560 | 0.4966 | 1.1437 | 1.1590 | 0.8335 | 0.5812 | 1.1786 | 0.7946 |
| B | -0.2903 | 0.0952 | -1.1494 | -1.8772 | -0.5827 | -1.0647 | 0.4136 | 1.4690 | 0.1840 | 0.4577 | 0.7038 | 1.2962 |
| C | 0.0 | 0.0821 | 0.1228 | 0.1228 | 0.1131 | 0.1228 | 0.2618 | 0.1229 | 0.1327 | 0.0052 | 0.1425 | 0.1827 |
| Ability Categories > 3. | + 0.12 | + 0.18 | + 0.01 | + 0.34 | + 0.01 | + 0.39 | + 0.05 | - 5.73 | + 0.19 | - 9.31 | + 0.09 | - 4.88 |
| 2.8 | + 0.34 | - 0.56 | + 0.02 | + 0.68 | + 0.03 | + 0.82 | + 0.18 | - 1.50 | - 0.44 | + 0.02 | - 1.38 | + 0.07 |
| 2.4 | - 0.34 | - 0.41 | - 9.07 | + 0.00 | + 0.15 | + 2.66 | - 0.01 | - 7.81 | + 2.16 | - 2.00 | + 0.26 | + 0.35 |
| 2.0 | - 2.70 | - 3.09 | - 0.52 | - 0.02 | - 9.70 | + 3.13 | - 1.50 | - 2.43 | + 0.76 | - 2.43 | + 0.01 | + 0.28 |
| 1.6 | - 0.09 | - 2.72 | - 2.64 | + 1.91 | - 1.29 | + 4.57 | + 0.26 | + 0.29 | + 1.03 | - 4.28 | - 0.65 | + 5.88 |
| 1.2 | - 1.81 | - 0.25 | - 2.64 | + 0.00 | - 2.33 | + 0.02 | - 0.03 | + 5.16 | + 0.01 | + 0.48 | - 0.23 | - 2.53 |
| .8 | - 0.06 | + 2.53 | - 6.62 | + 0.00 | - 0.03 | - 0.01 | - 0.20 | + 2.23 | - 0.05 | + 2.81 | + 0.05 | - 0.62 |
| .4 | - 0.27 | + 0.72 | - 0.04 | - 0.34 | + 0.71 | - 3.00 | + 1.11 | - 1.43 | - 1.59 | + 1.95 | + 1.18 | - 0.05 |
| .0 | + 3.25 | + 3.70 | + 0.13 | + 1.69 | + 0.24 | - 1.83 | - 0.38 | - 7.82 | - 0.25 | + 1.47 | - 0.05 | + 0.87 |
| -.4 | + 0.54 | - 1.53 | + 9.65 | - 2.36 | + 2.29 | - 0.27 | - 1.45 | - 4.23 | + 0.62 | - 0.04 | - 0.13 | - 0.15 |
| -.8 | + 0.46 | - 0.42 | + 0.30 | - 0.80 | - 0.34 | - 0.01 | + 1.27 | - 0.96 | + 0.45 | - 6.94 | - 4.21 | - 0.03 |
| -1.2 | - 1.39 | - 0.04 | - 0.01 | + 0.65 | - 2.10 | + 3.00 | + 0.59 | + 8.44 | + 0.01 | - 3.41 | + 2.27 | + 3.29 |
| -1.6 | - 1.99 | - 0.00 | - 8.16 | + 0.02 | - 0.65 | + 4.36 | + 3.30 | + 10.29 | + 0.33 | - 0.02 | + 0.15 | - 0.41 |
| -2.0 | - 0.87 | - 0.04 | - 1.45 | + 0.10 | - 0.00 | - 0.04 | - 0.78 | + 7.37 | + 0.08 | + 3.79 | + 2.33 | + 0.12 |
| -2.4 | - 0.72 | + 0.03 | - 0.38 | + 0.60 | + 0.05 | - 0.79 | - 0.02 | + 0.78 | - 0.42 | + 2.96 | + 4.68 | + 0.01 |
| -2.8 | - 0.27 | + 6.06 | + 0.15 | - 0.39 | + 1.41 | - 0.76 | + 4.09 | + 1.09 | - 1.17 | - 0.27 | + 0.12 | - 0.76 |
| <-3. | - 0.04 | + 4.79 | - 0.00 | + 0.20 | + 4.01 | - 0.75 | + 1.14 | + 1.47 | - 0.01 | - 0.70 | + 1.84 | + 2.25 |
| TOTAL CHI | 15.18 | 24.11 | 41.79 | 10.11 | 25.13 | 26.41 | 16.37 | 69.03 | 9.56 | 42.89 | 19.61 | 22.56 |

Table 7 (cont'd)

| Item Parameters | Item Numbers 25 | 26 | 27 | 28 | 29 | 30 | 31 * | 32 | 33 | 34 | 35 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.8915 | 0.9420 | 0.6449 | 0.7727 | 1.3525 | 1.1607 | 0.7028 | 0.5636 | 1.2336 | 0.8148 | 0.8458 | 0.7010 |
| B | 0.5877 | 0.7042 | 0.4812 | 1.0877 | 1.0347 | 1.6635 | -2.0581 | -1.1295 | -0.3188 | -0.5085 | -0.4746 | -0.1348 |
| C | 0.1547 | 0.2230 | 0.0251 | 0.1155 | 0.1055 | 0.1517 | 0.0673 | 0.0673 | 0.1344 | 0.0673 | 0.1631 | 0.1313 |
| Ability Categories | | | | | | | | | | | | |
| > 3. | + 0.18 | + 0.16 | + 0.54 | + 0.60 | + 0.07 | - 0.78 | + 0.03 | - 2.50 | + 0.01 | + 0.09 | + 0.07 | - 2.17 |
| 2.8 | - 0.18 | - 0.26 | - 0.08 | + 0.51 | - 1.16 | + 0.47 | - 12.39 | - 0.45 | + 0.03 | + 0.24 | - 1.99 | - 0.21 |
| 2.4 | - 0.43 | + 1.47 | - 3.05 | - 0.03 | - 0.31 | - 0.85 | + 0.26 | - 3.00 | - 2.32 | - 1.21 | + 0.77 | + 0.04 |
| 2.0 | - 1.12 | - 0.38 | - 0.14 | - 0.19 | + 0.09 | - 1.12 | - 13.90 | + 2.11 | + 0.00 | - 0.71 | - 1.21 | - 0.11 |
| 1.6 | - 0.48 | + 0.05 | + 0.02 | - 0.00 | + 0.09 | + 1.44 | - 0.20 | + 2.81 | - 0.36 | - 0.45 | - 0.09 | + 0.42 |
| 1.2 | + 6.94 | - 0.00 | + 0.26 | - 1.01 | + 0.49 | + 0.01 | - 2.28 | + 3.07 | - 3.52 | + 0.62 | + 2.11 | + 1.22 |
| .8 | + 1.26 | - 0.31 | - 0.63 | + 1.80 | + 0.09 | + 0.17 | - 0.19 | + 0.12 | + 0.02 | + 1.09 | - 0.16 | + 0.71 |
| .4 | - 0.74 | - 0.07 | + 0.93 | + 0.65 | - 3.73 | + 0.00 | - 0.16 | - 0.01 | + 0.03 | - 3.19 | + 1.75 | - 1.50 |
| .0 | - 3.95 | + 0.23 | - 0.10 | - 1.26 | + 0.71 | - 2.34 | + 0.23 | - 0.91 | + 2.55 | + 0.01 | - 0.51 | - 0.77 |
| -.4 | - 0.06 | - 0.05 | + 0.63 | - 0.04 | + 0.00 | - 1.51 | + 0.31 | - 0.22 | - 0.01 | + 1.79 | - 0.73 | - 0.10 |
| -.8 | + 1.78 | - 0.01 | - 0.09 | - 0.34 | + 1.17 | + 1.40 | + 4.53 | - 4.10 | - 2.19 | - 0.09 | + 0.17 | + 0.67 |
| -1.2 | + 2.42 | + 1.30 | - 1.14 | + 0.00 | + 1.75 | + 7.44 | + 1.78 | - 0.00 | - 0.11 | + 0.03 | - 0.66 | - 0.01 |
| -1.6 | + 0.08 | + 1.21 | - 0.00 | + 0.03 | - 1.78 | - 0.07 | - 1.51 | + 3.83 | + 1.54 | - 0.01 | - 0.00 | + 0.45 |
| -2.0 | + 0.08 | + 0.22 | - 0.00 | + 2.99 | + 0.41 | + 0.10 | - 8.81 | + 2.91 | - 0.15 | - 2.41 | + 7.29 | + 0.71 |
| -2.4 | + 0.26 | + 0.28 | + 0.33 | + 4.75 | - 0.04 | + 6.16 | - 0.07 | + 0.30 | + 1.72 | - 0.28 | + 0.02 | + 1.91 |
| -2.8 | - 0.09 | + 5.43 | + 0.13 | + 0.01 | + 3.24 | + 0.01 | + 0.03 | + 0.14 | + 1.09 | + 0.05 | + 0.99 | - 1.96 |
| <-3. | - 0.28 | + 0.14 | - 1.33 | + 0.12 | + 0.48 | - 0.14 | - 1.75 | - 0.13 | + 5.77 | + 0.01 | + 0.29 | + 0.00 |
| TOTAL CHI | 20.33 | 11.57 | 9.43 | 14.32 | 15.60 | 24.00 | 48.44 | 26.62 | 20.92 | 12.27 | 18.83 | 12.96 |

| Item Parameters | Item Numbers 37 | 38 | 39 * | 40 | 41 | 42 | 43 | 44 * | 45 | 46 * | 47 | 48 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.8192 | 0.8368 | 1.0499 | 0.4337 | 1.1560 | 0.4966 | 1.1437 | 1.1590 | 0.8335 | 0.5812 | 1.1786 | 0.7946 |
| B | -0.2903 | 0.0952 | -1.1494 | -1.8772 | -0.5827 | -1.0647 | 0.4136 | 1.4690 | 0.1840 | 0.4577 | 0.7038 | 1.2962 |
| C | 0.0 | 0.0821 | 0.1228 | 0.1228 | 0.1131 | 0.1228 | 0.2618 | 0.1229 | 0.1327 | 0.0052 | 0.1425 | 0.1827 |
| Ability Categories | | | | | | | | | | | | |
| > 3. | + 0.12 | + 0.18 | + 0.01 | + 0.34 | + 0.01 | + 0.39 | + 0.05 | - 5.73 | + 0.19 | - 9.31 | + 0.09 | - 4.89 |
| 2.8 | + 0.34 | - 0.56 | + 0.02 | + 0.68 | + 0.03 | + 0.82 | + 0.18 | - 1.50 | - 0.44 | + 0.02 | - 1.38 | + 0.07 |
| 2.4 | - 0.34 | - 0.41 | - 9.07 | + 0.00 | + 0.15 | + 2.66 | - 0.01 | - 7.81 | + 2.16 | - 2.00 | + 0.26 | + 0.35 |
| 2.0 | - 2.70 | - 3.09 | - 0.52 | - 0.02 | - 9.70 | + 3.13 | - 1.50 | - 2.43 | + 0.76 | - 2.43 | + 0.01 | + 0.28 |
| 1.6 | - 0.00 | - 2.72 | - 2.64 | + 1.91 | - 1.29 | + 4.57 | + 0.26 | + 0.29 | + 1.03 | - 4.28 | - 0.65 | + 5.88 |
| 1.2 | - 1.81 | - 0.25 | - 2.64 | + 0.00 | - 2.33 | + 0.02 | - 0.03 | + 5.16 | + 0.01 | + 0.48 | - 0.23 | - 2.53 |
| .8 | - 0.06 | + 2.53 | - 6.62 | + 0.00 | - 0.03 | - 0.01 | - 0.20 | + 2.23 | - 0.05 | + 2.81 | + 0.05 | - 0.62 |
| .4 | - 0.27 | + 0.72 | - 0.04 | - 0.34 | + 0.71 | - 3.00 | + 1.11 | - 1.43 | - 1.59 | + 1.95 | + 1.18 | - 0.05 |
| .0 | + 3.25 | + 3.70 | + 0.13 | + 1.69 | + 0.04 | - 1.83 | - 0.38 | - 7.82 | - 0.25 | + 1.47 | - 0.05 | + 0.87 |
| -.4 | + 0.54 | - 1.53 | + 9.65 | - 2.36 | + 2.29 | - 0.27 | - 1.45 | - 4.23 | + 0.62 | - 0.04 | - 0.13 | - 0.15 |
| -.8 | + 0.46 | - 0.42 | + 0.30 | - 0.80 | - 0.34 | - 0.01 | + 1.27 | - 0.96 | + 0.45 | - 6.94 | - 4.21 | - 0.03 |
| -1.2 | - 1.38 | - 0.04 | - 0.01 | + 0.65 | - 2.10 | + 3.00 | + 0.59 | + 8.44 | + 0.01 | - 3.41 | + 2.27 | + 3.29 |
| -1.6 | - 1.99 | - 0.00 | - 8.16 | + 0.02 | - 0.65 | + 4.36 | + 3.30 | + 10.29 | + 0.33 | - 0.02 | + 0.15 | - 0.41 |
| -2.0 | - 0.87 | - 0.04 | - 1.45 | + 0.10 | - 0.00 | - 0.04 | - 0.78 | + 7.37 | + 0.08 | + 3.79 | + 2.33 | + 0.12 |
| -2.4 | - 0.72 | + 0.03 | - 0.38 | + 0.60 | + 0.05 | - 0.79 | - 0.02 | + 0.78 | - 0.42 | + 2.96 | + 4.68 | + 0.01 |
| -2.8 | - 0.27 | + 6.06 | + 0.15 | - 0.39 | + 1.41 | - 0.76 | + 4.09 | + 1.09 | - 1.17 | - 0.27 | + 0.12 | - 0.76 |
| <-3. | - 0.04 | + 4.79 | - 0.00 | + 0.20 | + 4.01 | - 0.75 | + 1.14 | + 1.47 | - 0.01 | - 0.70 | + 1.84 | + 2.25 |
| TOTAL CHI | 15.18 | 24.11 | 41.79 | 10.11 | 25.13 | 26.41 | 16.37 | 69.03 | 9.56 | 42.89 | 19.61 | 22.56 |

Table 7 (cont'd)

| | Item Parameters | Item Numbers 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | 0.7104 | 1.6240 | 0.7483 | 0.6558 | 0.8215 | 0.7496 | 0.7927 | 1.0831 | 0.9281 | 1.1047 | 0.9603 | 0.6932 |
| | B | 1.6595 | 2.1587 | -1.9031 | -1.3465 | -0.5892 | -0.0724 | 1.6959 | 0.1668 | -0.3775 | -0.1436 | 1.6202 | -0.2895 |
| | C | 0.0913 | 0.1051 | 0.0673 | 0.1228 | 0.1228 | 0.0 | 0.1517 | 0.4326 | 0.0133 | 0.1052 | 0.1934 | 0.0673 |
| Ability Categories | > 3. | - 1.88 | - 0.10 | + 0.02 | - 10.67 | + 0.06 | + 0.22 | + 0.14 | + 0.03 | + 0.05 | + 0.02 | - 0.06 | + 0.22 |
| | 2.8 | - 3.25 | - 0.04 | + 0.05 | + 0.17 | + 0.15 | + 0.51 | - 0.33 | + 0.09 | + 0.13 | + 0.07 | + 0.00 | + 0.46 |
| | 2.4 | + 0.37 | - 1.42 | + 0.17 | + 0.60 | + 0.61 | - 0.77 | + 0.03 | - 0.81 | - 0.37 | + 0.35 | - 0.28 | + 1.69 |
| | 2.0 | + 1.20 | + 0.95 | - 0.34 | - 0.85 | - 1.99 | - 7.11 | - 0.03 | + 0.29 | + 2.06 | - 0.24 | + 0.29 | - 0.99 |
| | 1.6 | + 0.02 | + 0.50 | + 0.14 | - 3.00 | - 0.09 | - 1.77 | + 0.66 | + 0.07 | - 2.17 | - 0.01 | + 0.00 | - 0.00 |
| | 1.2 | + 0.02 | - 0.16 | + 0.01 | - 1.46 | - 0.36 | - 0.34 | - 0.36 | - 0.01 | + 1.78 | - 0.00 | + 0.08 | + 0.67 |
| | .8 | - 0.90 | - 0.80 | - 3.36 | + 0.01 | - 0.77 | + 0.86 | - 1.07 | + 0.01 | + 0.19 | + 0.22 | - 0.50 | - 1.34 |
| | .4 | + 0.84 | - 4.58 | - 1.26 | + 0.81 | - 1.80 | + 1.59 | + 1.02 | - 0.14 | - 0.99 | - 0.16 | - 0.00 | + 0.19 |
| | .0 | - 0.16 | + 0.34 | + 0.36 | + 1.35 | + 0.54 | + 3.53 | + 0.31 | - 0.01 | + 0.24 | + 0.21 | + 0.09 | - 0.14 |
| | -.4 | + 0.11 | + 0.12 | + 0.40 | + 0.41 | + 7.96 | - 2.67 | - 0.19 | - 0.01 | - 1.06 | - 0.11 | - 0.07 | + 0.48 |
| | -.8 | - 0.83 | + 1.61 | + 0.97 | - 0.77 | + 0.00 | - 2.56 | - 0.01 | - 0.08 | + 0.31 | - 0.52 | + 1.56 | - 0.63 |
| | -1.2 | - 0.11 | + 0.14 | + 0.29 | + 0.16 | - 2.18 | - 0.01 | - 0.44 | + 1.27 | + 0.05 | + 1.11 | + 1.41 | + 0.87 |
| | -1.6 | - 0.09 | + 1.03 | + 0.17 | - 0.51 | - 0.73 | + 0.72 | - 0.13 | + 2.39 | + 0.68 | + 0.08 | + 0.18 | - 0.00 |
| | -2.0 | + 2.59 | + 6.00 | - 2.52 | - 0.65 | - 4.06 | + 1.67 | + 0.27 | + 1.64 | + 0.21 | - 0.00 | - 0.04 | - 0.35 |
| | -2.4 | + 2.51 | + 0.02 | - 1.81 | - 0.11 | - 0.39 | - 0.19 | + 3.21 | + 0.03 | - 1.32 | + 0.42 | - 5.57 | - 0.14 |
| | -2.8 | + 0.00 | - 0.28 | - 0.34 | - 1.80 | - 1.11 | - 0.35 | + 0.46 | - 0.06 | - 0.09 | + 0.13 | + 2.02 | + 0.00 |
| | <-3. | + 0.19 | - 0.35 | - 0.12 | + 0.11 | + 0.60 | - 0.09 | - 0.12 | + 0.00 | - 0.17 | + 1.27 | + 0.05 | - 0.82 |
| | TOTAL CHI | 15.09 | 18.47 | 12.34 | 23.44 | 23.39 | 24.95 | 8.76 | 6.94 | 11.86 | 4.92 | 12.22 | 9.00 |

| | Item Parameters | Item Numbers 61 | 62 | 63 | 64 * | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | 0.8078 | 0.5542 | 1.0405 | 0.4447 | 1.2287 | 0.9303 | 0.7261 | 0.7850 | 1.2901 | 0.7770 | 0.8847 | 0.9985 |
| | B | 0.9036 | -1.7510 | -0.8634 | -0.6354 | -0.0991 | -0.1497 | 0.2010 | 0.1460 | -1.5460 | -1.2709 | -0.6691 | 0.0016 |
| | C | 0.1661 | 0.0673 | 0.0660 | 0.0673 | 0.0959 | 0.1510 | 0.1195 | 0.1999 | 0.1228 | 0.1228 | 0.1228 | 0.0867 |
| Ability Categories | > 3. | + 0.47 | - 3.39 | + 0.01 | + 0.01 | + 0.02 | + 0.16 | + 0.47 | - 0.92 | + 0.00 | + 0.05 | + 0.06 | + 0.08 |
| | 2.8 | - 8.39 | - 0.72 | + 0.05 | + 2.40 | + 0.08 | - 0.64 | + 1.25 | - 0.05 | + 0.00 | - 5.18 | - 4.13 | - 1.85 |
| | 2.4 | + 1.95 | - 0.17 | - 2.33 | + 7.61 | + 0.42 | + 1.93 | - 0.05 | + 0.42 | + 0.01 | - 3.65 | + 0.73 | - 0.39 |
| | 2.0 | + 1.63 | - 0.07 | + 0.00 | + 3.60 | - 0.00 | + 0.08 | + 0.46 | + 0.11 | + 0.07 | + 0.49 | + 1.08 | - 1.59 |
| | 1.6 | - 0.32 | - 0.53 | - 5.72 | + 1.79 | + 0.05 | + 1.55 | - 0.31 | - 2.27 | - 2.07 | - 2.30 | - 0.07 | - 3.87 |
| | 1.2 | + 0.85 | - 0.46 | - 0.26 | + 1.62 | + 0.63 | - 0.10 | - 0.02 | + 0.98 | - 24.50 | - 0.73 | - 1.12 | - 0.42 |
| | .8 | - 2.04 | - 6.44 | + 0.21 | - 0.34 | - 0.25 | - 1.07 | - 0.03 | + 2.28 | - 0.01 | + 0.07 | - 2.23 | + 2.53 |
| | .4 | + 0.78 | + 0.09 | - 0.01 | - 2.48 | - 0.00 | + 0.34 | - 0.04 | + 0.04 | - 0.21 | + 0.04 | + 1.12 | + 0.01 |
| | .0 | + 1.37 | + 0.18 | - 0.41 | - 1.56 | - 0.09 | - 0.39 | + 0.09 | - 2.11 | - 0.45 | + 0.01 | + 0.69 | + 0.93 |
| | -.4 | - 4.61 | + 4.98 | + 5.04 | - 0.70 | + 0.03 | + 0.77 | - 0.03 | + 0.01 | + 0.82 | + 1.14 | - 1.12 | + 0.49 |
| | -.8 | - 0.06 | + 1.71 | - 0.26 | - 2.06 | - 0.00 | - 0.67 | + 0.88 | - 0.69 | + 1.10 | + 0.01 | + 3.47 | - 5.29 |
| | -1.2 | + 1.93 | - 0.20 | - 0.40 | + 4.45 | + 0.00 | + 0.13 | - 0.50 | + 2.35 | + 0.66 | - 0.04 | - 2.51 | - 2.23 |
| | -1.6 | + 0.89 | - 0.04 | - 1.72 | + 2.36 | + 0.82 | - 0.02 | - 0.53 | - 0.08 | - 1.46 | - 0.39 | + 0.01 | + 1.24 |
| | -2.0 | + 0.59 | - 1.05 | - 0.00 | + 3.25 | + 0.01 | + 2.39 | + 0.43 | - 0.22 | - 0.03 | - 0.38 | - 1.13 | + 0.92 |
| | -2.4 | + 0.54 | - 1.78 | + 0.42 | - 0.03 | + 0.75 | + 0.52 | + 1.99 | + 2.00 | - 3.62 | - 0.42 | + 0.09 | + 0.11 |
| | -2.8 | - 0.36 | - 4.31 | + 7.52 | + 0.11 | + 2.85 | - 0.95 | - 0.14 | + 1.76 | - 0.22 | + 0.50 | + 0.50 | + 2.02 |
| | <-3. | - 1.81 | - 0.54 | + 0.01 | + 0.61 | + 0.00 | + 1.56 | + 1.03 | + 3.56 | + 7.67 | - 0.09 | - 0.29 | + 4.14 |
| | TOTAL CHI | 28.18 | 26.37 | 24.38 | 34.96 | 6.00 | 13.56 | 8.25 | 19.86 | 42.90 | 15.43 | 20.35 | 28.11 |

Table 7 (cont'd)

| Item Parameters | | Item Numbers | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 73 * | 74 * | 75 | 76 * | 77 | 78 | 79 | 80 | 81 | 82 * | 83 | 84 |
| | A | 0.6759 | 1.3059 | 0.9767 | 0.4008 | 0.4974 | 0.8694 | 0.8990 | 1.2773 | 0.7220 | 0.6362 | 0.3937 | 1.2484 |
| | B | 1.0616 | 1.0332 | 1.9264 | -0.4242 | -0.5061 | 0.4082 | 0.4073 | 1.3546 | 2.3811 | -1.3637 | -1.9519 | -0.7481 |
| | C | 0.2148 | 0.0505 | 0.1001 | 0.0673 | 0.0673 | 0.0737 | 0.1701 | 0.1308 | 0.1367 | 0.0673 | 0.0673 | 0.0933 |
| Ability Categories | > 3. | - 0.32 | + 0.16 | + 0.59 | + 0.29 | - 0.02 | - 1.52 | + 0.25 | - 8.89 | + 0.31 | + 0.18 | - 0.04 | + 0.01 |
| | 2.8 | - 2.05 | - 2.63 | + 0.84 | - 0.21 | + 0.42 | + 0.96 | - 1.53 | - 0.21 | + 0.65 | - 1.05 | + 0.09 | + 0.02 |
| | 2.4 | + 0.48 | - 0.43 | - 0.00 | - 0.65 | + 0.18 | - 1.23 | - 2.67 | + 0.21 | + 0.84 | + 1.32 | - 1.13 | - 8.23 |
| | 2.0 | - 5.52 | - 2.40 | - 1.25 | - 0.04 | + 0.21 | + 1.85 | + 0.05 | - 0.03 | - 2.74 | + 2.66 | - 0.05 | + 0.51 |
| | 1.6 | + 2.02 | + 1.78 | + 0.23 | + 0.23 | + 3.92 | - 0.72 | - 0.28 | + 0.00 | + 0.00 | + 1.12 | - 2.98 | - 0.74 |
| | 1.2 | + 0.74 | - 0.74 | + 0.01 | + 2.95 | + 0.29 | + 0.00 | + 0.25 | + 0.09 | - 0.26 | + 2.19 | - 2.56 | + 0.16 |
| | .8 | + 4.50 | + 0.04 | - 0.91 | + 2.75 | + 1.04 | + 0.09 | - 0.85 | + 0.9[illegible] | + 1.97 | + 3.88 | - 2.91 | - 0.02 |
| | .4 | - 0.72 | + 0.35 | + 3.66 | - 0.17 | - 1.07 | - 0.34 | + 1.55 | - 0.39 | - 0.00 | - 0.49 | + 3.92 | + 0.63 |
| | .0 | - 3.91 | + 0.59 | - 0.85 | + 0.01 | - 3.85 | - 0.91 | + 0.00 | - 0.63 | + 0.03 | - 1.89 | + 1.62 | - 1.28 |
| | -.4 | - 0.87 | - 1.86 | - 0.43 | - 4.83 | - 1.00 | + 0.17 | - 0.00 | - 3.88 | - 0.41 | - 1.93 | + 3.41 | + 0.49 |
| | -.8 | + 1.33 | - 1.55 | + 0.56 | - 5.94 | + 1.09 | - 0.10 | - 0.97 | + 0.58 | - 3.59 | - 0.81 | - 0.25 | + 0.07 |
| | -1.2 | + 6.39 | - 0.31 | + 1.95 | + 0.00 | - 0.22 | + 0.04 | - 0.06 | + 1.53 | + 1.75 | - 0.30 | - 0.29 | - 0.31 |
| | -1.6 | - 1.20 | + 0.23 | - 0.38 | + 3.72 | + 3.06 | + 0.09 | + 0.28 | + 4.03 | + 0.45 | + 0.12 | + 0.21 | - 0.02 |
| | -2.0 | + 2.90 | + 0.23 | - 0.41 | + 4.64 | + 0.61 | - 0.07 | + 1.84 | + 4.90 | + 1.74 | + 9.49 | - 1.50 | + 0.12 |
| | -2.4 | + 0.01 | + 9.38 | + 0.82 | + 6.10 | + 2.03 | + 0.78 | + 2.98 | + 2.21 | - 0.02 | + 4.65 | - 9.06 | - 0.41 |
| | -2.8 | - 0.80 | + 7.70 | - 0.07 | - 0.05 | - 0.46 | + 1.51 | - 0.06 | + 0.74 | + 1.12 | + 5.16 | - 0.12 | + 0.35 |
| | <-3. | + 1.51 | + 4.76 | + 0.07 | + [illegible] | + 0.31 | + 1.12 | + 8.89 | + 0.08 | + 0.35 | + 3.85 | - 0.08 | + 4.09 |
| TOTAL CHI | | 34.25 | 35.15 | 13.04 | 36.68 | 19.79 | 9.80 | 22.53 | 29.31 | 16.23 | 41.08 | 30.22 | 17.46 |

| Item Parameters | | Item Numbers | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 85 | 86 | 87 | 88 * | 89 | 90 * | 91 | 92 | 93 | 94 | 95 | 96 |
| | A | 1.1055 | 1.2679 | 1.0439 | 0.4713 | 1.0227 | 0.8396 | 1.1620 | 0.6454 | 0.6285 | 1.1100 | 0.2828 | 0.8676 |
| | B | -0.1603 | 0.4799 | 0.0428 | -0.3102 | -1.4264 | -1.1618 | -0.2679 | 0.1979 | 0.5859 | 0.4262 | -1.7099 | 0.6004 |
| | C | 0.1425 | 0.2593 | 0.1709 | 0.0673 | 0.1228 | 0.1228 | 0.2593 | 0.2053 | 0.1839 | 0.1690 | 0.0673 | 0.1702 |
| Ability Categories | > 3. | + 0.04 | + 0.05 | + 0.08 | - 1.24 | + 0.01 | + 0.05 | + 0.02 | + 0.79 | + 1.34 | + 0.11 | + 0.83 | + 0.47 |
| | 2.8 | + 0.13 | + 0.20 | + 0.24 | - 0.84 | + 0.02 | + 0.12 | + 0.07 | - 0.11 | + 0.14 | + 0.36 | + 0.06 | - 0.61 |
| | 2.4 | - 3.36 | - 3.22 | + 0.00 | + 0.10 | + 0.10 | - 5.28 | + 0.33 | - 0.50 | - 0.82 | - 3.29 | - 2.72 | + 0.08 |
| | 2.0 | + 1.14 | - 0.66 | - 0.37 | - 3.90 | + 0.43 | - 9.08 | - 1.28 | + 3.64 | - 0.13 | - 0.73 | - 0.21 | - 3.26 |
| | 1.6 | - 0.34 | + 0.04 | - 2.50 | - 5.34 | - 0.34 | - 10.89 | + 0.64 | - 0.15 | + 0.13 | + 0.04 | + 0.72 | + 0.76 |
| | 1.2 | + 0.26 | - 0.68 | + 0.00 | - 4.50 | - 0.57 | - 6.78 | + 4.14 | - 0.10 | - 0.03 | + 0.65 | + 1.75 | - 1.92 |
| | .8 | + 0.05 | + 0.25 | + 1.44 | - 0.00 | - 3.32 | - 10.42 | - 0.37 | - 0.00 | - 0.15 | - 0.00 | + 0.14 | + 1.27 |
| | .4 | + 0.91 | + 0.22 | - 0.03 | + 6.57 | + 2.21 | + 0.06 | - 0.02 | - 0.08 | + 0.85 | + 1.01 | - 1.75 | + 6.24 |
| | .0 | - 1.49 | + 1.12 | + 0.43 | + 10.94 | + 1.32 | + 7.64 | + 0.23 | - 0.18 | - 0.00 | - 1.43 | + 0.02 | - 0.64 |
| | -.4 | - 1.28 | - 3.91 | - 2.93 | + 0.78 | + 0.01 | + 10.28 | - 2.90 | + 0.59 | - 2.47 | - 0.56 | - 0.10 | - 7.60 |
| | -.8 | + 1.75 | - 0.16 | - 1.07 | - 3.56 | - 0.47 | + 9.30 | + 0.73 | - 0.40 | - 0.00 | + 0.11 | - 0.10 | - 0.72 |
| | -1.2 | + 0.00 | + 0.93 | + 7.54 | - 1.05 | - 0.62 | - 3.16 | + 1.25 | + 0.00 | + 3.56 | + 0.68 | + 0.03 | + 4.16 |
| | -1.6 | + 1.75 | + 0.71 | + 1.59 | - 8.01 | + 0.34 | - 24.21 | - 0.09 | + 0.22 | + 1.31 | + 2.63 | + 0.16 | + 2.93 |
| | -2.0 | - 0.40 | + 10.14 | + 1.12 | - 0.10 | - 0.02 | - 4.96 | + 0.85 | + 0.75 | - 0.82 | + 0.08 | + 0.80 | + 2.59 |
| | -2.4 | + 9.35 | + 2.91 | + 0.01 | - 0.60 | - 0.09 | - 4.07 | - 0.01 | + 0.31 | + 0.12 | + 0.27 | + 0.00 | - 0.18 |
| | -2.8 | - 0.00 | + 1.66 | - 0.11 | - 0.53 | + 1.62 | - 0.39 | + 6.69 | - 1.77 | + 0.11 | - 0.58 | + 0.26 | + 0.77 |
| | <-3. | - 2.03 | + 0.85 | + 0.34 | + 2.10 | + 4.59 | + 0.36 | - 0.05 | + 2.37 | - 0.91 | - 0.09 | - 1.46 | - 0.25 |
| TOTAL CHI | | 24.39 | 27.21 | 19.79 | 50.17 | 16.07 | 107.04 | 19.67 | 11.95 | 12.91 | 12.63 | 11.12 | 34.45 |

Table 7 (cont'd)

| Item Parameters | 97 | 98 | 99 | 100 | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.9037 | 0.7079 | 0.5323 | 1.2784 | 0.4779 | 0.9738 | 0.5300 | 0.8343 | 1.3571 | 0.6919 | 1.1251 | 0.4250 |
| B | 0.7167 | 1.0162 | -1.1351 | 1.9422 | -2.3850 | -0.9235 | -0.8485 | -0.3070 | 1.7021 | 2.3840 | 1.2667 | -0.8533 |
| C | 0.1008 | 0.1001 | 0.0111 | 0.1134 | 0.0673 | 0.0673 | 0.0673 | 0.0026 | 0.0908 | 0.0292 | 0.1283 | 0.0673 |
| **Ability Categories** | | | | | | | | | | | | |
| > 3. | + 0.51 | - 7.32 | - 2.61 | - 4.35 | + 0.07 | + 0.01 | + 0.18 | + 0.05 | + 0.21 | + 0.64 | + 0.16 | - 1.24 |
| 2.8 | + 1.38 | - 0.03 | - 0.23 | - 0.09 | + 0.21 | - 30.75 | + 0.53 | + 0.19 | + 1.15 | - 0.04 | - 0.22 | - 0.00 |
| 2.4 | - 4.44 | + 0.12 | + 0.78 | - 0.23 | + 0.72 | + 0.15 | - 0.03 | - 0.05 | - 1.03 | + 0.17 | + 0.03 | - 0.32 |
| 2.0 | - 0.59 | + 3.02 | - 8.76 | + 2.91 | - 4.02 | - 3.09 | + 0.05 | + 0.38 | - 0.00 | - 0.27 | - 0.85 | - 8.60 |
| 1.6 | + 1.32 | - 0.28 | + 1.48 | + 0.00 | + 0.27 | - 0.38 | - 2.39 | - 1.25 | + 0.00 | - 0.01 | + 0.51 | - 0.00 |
| 1.2 | - 0.00 | + 0.49 | + 0.71 | - 0.10 | - 0.00 | + 0.00 | - 1.43 | - 0.03 | + 0.86 | - 0.01 | + 0.80 | + 0.25 |
| .8 | + 0.05 | + 0.30 | + 0.47 | + 0.22 | - 1.12 | + 0.05 | + 0.38 | + 1.08 | - 0.01 | + 0.33 | - 0.62 | + 0.62 |
| .4 | - 0.09 | - 2.34 | + 0.02 | - 1.01 | - 6.32 | - 1.11 | + 2.06 | - 4.07 | - 1.99 | + 0.04 | - 0.34 | - 0.67 |
| .0 | + 0.05 | - 1.33 | - 0.17 | - 2.72 | + 0.35 | - 0.03 | + 0.00 | + 0.52 | + 0.31 | - 1.17 | + 1.81 | + 0.01 |
| -.4 | + 0.01 | + 2.51 | + 1.40 | + 0.02 | + 3.17 | + 2.92 | - 0.75 | + 1.89 | - 0.09 | + 0.75 | - 0.30 | + 1.19 |
| -.8 | - 0.70 | + 0.43 | - 3.28 | + 2.35 | + 4.14 | + 0.03 | - 0.87 | - 0.03 | + 3.28 | + 0.12 | - 3.31 | + 4.14 |
| -1.2 | - 0.79 | - 0.22 | - 0.53 | + 6.66 | + 0.15 | + 0.01 | + 5.40 | + 0.22 | - 0.46 | - 1.03 | + 5.97 | - 0.34 |
| -1.6 | + 4.84 | - 0.27 | - 0.01 | + 0.05 | - 3.82 | - 0.81 | + 0.24 | - 5.75 | - 1.27 | + 0.49 | - 0.34 | - 4.33 |
| -2.0 | - 0.15 | + 3.54 | - 0.17 | - 0.88 | - 0.19 | - 2.36 | - 0.56 | - 0.36 | - 0.00 | - 0.42 | + 0.07 | - 3.13 |
| -2.4 | + 1.13 | - 2.35 | + 0.69 | + 0.92 | - 0.26 | + 0.34 | - 5.15 | + 2.27 | + 2.22 | + 0.04 | + 7.38 | - 0.15 |
| -2.8 | - 0.02 | + 0.01 | + 0.30 | + 1.09 | - 0.28 | - 2.22 | - 0.10 | - 0.26 | + 1.95 | + 0.51 | + 0.92 | + 0.59 |
| <-3. | + 0.61 | - 0.01 | + 1.57 | - 0.07 | - 1.70 | - 0.69 | - 0.00 | + 0.28 | + 5.81 | - 0.19 | - 0.51 | - 1.15 |
| TOTAL CHI | 16.68 | 24.57 | 23.18 | 23.67 | 26.79 | [illegible].94 | 20.12 | 18.69 | 20.64 | 6.22 | 24.13 | 26.74 |

| Item Parameters | 109 | 110* | 111 | 112 | 113 | 114* | 115 | 116 | 117 | 118 | 119 | 120 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1.0282 | 0.5737 | 0.2729 | 1.5804 | 1.1986 | 1.7500 | 1.7500 | 0.8021 | 1.1056 | 0.6412 | 1.3872 | 1.3113 |
| B | 0.1542 | -0.4150 | -0.9318 | 1.4062 | 1.5804 | 2.3176 | 2.3166 | 0.7871 | 2.0503 | 2.6188 | 2.0110 | 1.2148 |
| C | 0.2219 | 0.0673 | 0.1228 | 0.1745 | 0.1695 | 0.1443 | 0.0893 | 0.1195 | 0.1474 | 0.1359 | 0.0865 | 0.0913 |
| **Ability Categories** | | | | | | | | | | | | |
| > 3. | + 0.03 | + 0.20 | - 1.70 | + 0.05 | + 0.22 | - 0.75 | - 0.18 | - 2.64 | - 0.22 | - 1.20 | + 0.41 | + 0.07 |
| 2.8 | + 0.12 | + 0.62 | + 2.30 | + 0.32 | - 4.19 | - 8.67 | + 0.29 | - 1.78 | + 2.71 | + 1.19 | - 5.00 | + 0.24 |
| 2.4 | + 0.60 | + 0.64 | + 0.12 | - 0.35 | - 4.87 | + 2.53 | - 0.00 | - 0.06 | - 0.93 | - 0.88 | - 0.18 | + 0.01 |
| 2.0 | - 0.88 | + 3.58 | - 0.47 | - 1.49 | + 1.21 | + 3.84 | + 0.12 | - 0.01 | - 0.00 | + 2.96 | + 0.73 | - 0.06 |
| 1.6 | - 4.98 | + 4.13 | - 0.05 | - 0.00 | + 0.18 | - 3.08 | - 0.09 | + 0.46 | - 0.00 | - 1.73 | + 0.27 | - 0.25 |
| 1.2 | - 0.23 | + 2.24 | - 3.53 | + 1.65 | + 0.79 | + 0.01 | + 0.07 | + 0.07 | + 0.26 | + 0.08 | + 0.59 | + 0.20 |
| .8 | + 0.08 | - 0.29 | - 0.09 | - 0.52 | - 0.15 | - 2.81 | - 1.67 | - 0.14 | - 0.07 | - 0.02 | - 4.38 | + 0.83 |
| .4 | + 5.96 | + 0.17 | + 0.28 | + 0.14 | - 0.86 | - 5.27 | + 0.51 | + 0.39 | + 0.59 | + 0.62 | + 0.15 | - 1.06 |
| .0 | + 0.00 | - 8.57 | + 0.36 | - 0.29 | - 0.28 | - 10.53 | - 0.01 | + 0.18 | - 1.03 | + 0.01 | - 0.50 | - 0.59 |
| -.4 | - 5.54 | - 1.21 | + 0.79 | - 1.35 | - 0.01 | - 0.91 | - 1.76 | - 1.36 | - 0.50 | - 2.10 | + 0.22 | + 1.74 |
| -.8 | - 0.06 | - 1.91 | + 0.97 | - 0.42 | + 1.17 | + 0.42 | - 0.26 | - 0.43 | + 0.23 | + 0.15 | - 0.05 | + 0.05 |
| -1.2 | + 1.20 | + 5.93 | - 1.05 | + 3.81 | - 0.32 | + 16.45 | + 0.18 | + 0.17 | + 0.00 | - 0.38 | + 0.20 | + 0.84 |
| -1.6 | + 0.11 | + 1.28 | + 1.11 | + 0.02 | + 0.32 | + 28.92 | + 1.23 | + 1.05 | + 0.01 | + 0.27 | + 0.37 | - 0.95 |
| -2.0 | + 5.23 | + 9.40 | - 0.89 | + 1.23 | + 15.35 | + 5.38 | + 6.02 | - 0.09 | + 2.10 | + 7.30 | - 0.00 | + 1.45 |
| -2.4 | + 0.06 | + 0.78 | - 3.67 | + 4.29 | + 0.03 | + 1.20 | + 1.78 | + 0.40 | + 4.09 | - 0.09 | + 6.44 | - 0.72 |
| -2.8 | + 1.17 | + 3.00 | - 0.88 | + 1.00 | - 0.12 | + 2.08 | + 3.13 | + 3.20 | + 0.56 | + 0.36 | + 0.07 | - 0.03 |
| <-3. | + 2.69 | - 1.44 | - 1.13 | + 2.13 | - 0.17 | - 0.24 | - 0.19 | + 0.00 | + 0.18 | + 0.01 | + 0.00 | + 1.42 |
| TOTAL CHI | 28.96 | 44.37 | 19.39 | 19.06 | 30.23 | 92.99 | 17.51 | 12.44 | 13.47 | 19.34 | 19.55 | 10.51 |

Table 7 (cont'd)

| Item Parameters | 121 | 122 | 123* | 124 | 125 | 126 | 127 | 128* | 129 | 130* | 131 | 132 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.6432 | 0.9720 | 0.5208 | 1.3795 | 0.9550 | 1.2157 | 0.9842 | 1.7500 | 0.9743 | 1.3966 | 0.6162 | 0.9537 |
| B | -2.2698 | 0.5815 | -0.0772 | -0.0218 | 1.8790 | 1.2520 | 1.8806 | 2.3687 | -1.4698 | -0.5012 | -0.9196 | 0.4231 |
| C | 0.0673 | 0.3348 | 0.0673 | 0.1691 | 0.2101 | 0.1315 | 0.1262 | 0.0716 | 0.1228 | 0.0809 | 0.1228 | 0.1307 |
| **Ability Categories** | | | | | | | | | | | | |
| > 3. | + 0.03 | + 0.11 | + 0.65 | + 0.01 | + 1.08 | + 0.15 | + 0.00 | + 0.89 | + 0.00 | + 0.00 | + 0.15 | + 0.12 |
| 2.8 | + 0.06 | - 1.70 | + 1.23 | + 0.02 | + 1.10 | + 0.65 | - 0.00 | + 2.20 | + 0.01 | + 0.01 | + 0.30 | - 8.91 |
| 2.4 | + 0.22 | + 0.05 | + 2.08 | + 0.12 | + 0.01 | + 0.42 | - 0.07 | + 0.01 | + 0.06 | + 0.04 | - 0.98 | - 0.35 |
| 2.0 | - 0.03 | + 4.13 | - 0.25 | + 0.74 | + 0.80 | - 0.37 | - 0.15 | - 0.03 | + 0.28 | - 2.30 | + 2.12 | - 2.62 |
| 1.6 | - 4.50 | - 0.08 | + 1.73 | + 0.03 | - 3.21 | - 0.31 | + 0.09 | - 0.00 | + 0.84 | + 0.06 | - 0.01 | + 0.01 |
| 1.2 | + 0.03 | - 2.74 | + 0.33 | - 0.01 | + 0.22 | + 0.23 | - 0.34 | - 2.18 | + 0.17 | - 3.54 | - 0.57 | + 0.08 |
| .8 | - 0.37 | + 0.46 | - 0.03 | + 0.73 | - 0.37 | + 0.28 | + 1.59 | - 6.45 | + 0.00 | - 1.35 | + 1.09 | + 0.47 |
| .4 | - 1.50 | + 0.63 | + 0.02 | - 0.72 | - 0.26 | - 0.37 | - 0.04 | - 6.77 | + 0.79 | + 0.00 | - 0.03 | + 0.58 |
| .0 | - 0.43 | - 0.98 | - 2.99 | + 0.00 | + 5.19 | - 0.63 | - 0.30 | + 0.07 | - 1.75 | + 0.07 | + 0.28 | + 0.01 |
| -.4 | + 5.49 | - 1.10 | - 0.37 | - 0.01 | + 0.33 | + 1.71 | - 1.11 | - 0.30 | - 1.47 | + 9.17 | - 2.19 | - 0.18 |
| -.8 | + 0.37 | + 0.92 | - 0.02 | + 1.09 | + 3.25 | + 0.87 | + 1.10 | + 11.82 | + 3.09 | - 1.02 | - 0.28 | - 4.01 |
| -1.2 | + 1.61 | + 5.84 | - 0.08 | + 0.05 | - 0.46 | + 0.00 | - 1.99 | + 1.37 | + 2.81 | - 7.04 | - 0.02 | - 0.14 |
| -1.6 | - 0.41 | + 0.07 | + 0.28 | + 0.05 | - 4.13 | + 0.83 | + 2.87 | + 0.23 | + 0.10 | - 2.98 | + 6.44 | + 2.23 |
| -2.0 | - 0.66 | + 0.03 | + 8.03 | - 2.13 | - 0.11 | - 1.82 | + 0.26 | + 1.29 | - 0.72 | + 0.66 | - 0.16 | + 0.23 |
| -2.4 | - 0.46 | + 0.28 | + 5.81 | + 0.03 | + 0.05 | - 0.29 | + 2.68 | + 2.81 | - 0.51 | - 0.12 | - 1.32 | + 0.23 |
| -2.8 | - 4.76 | - 0.14 | + 10.63 | + 0.62 | - 1.66 | - 1.06 | + 0.13 | + 4.83 | - 0.36 | + 8.65 | + 0.02 | + 3.92 |
| <-3. | - 1.41 | - 0.00 | - 0.04 | + 3.35 | - 2.67 | - 0.08 | + 3.80 | + 0.01 | - 0.72 | + 2.81 | + 0.07 | + 0.95 |
| TOTAL CHI | 22.35 | 19.25 | 34.57 | 9.20 | 24.86 | 10.11 | 16.53 | 41.27 | 10.67 | 39.83 | 16.04 | 25.03 |

| Item Parameters | 133 | 134 | 135 | 136* | 137 | 138 | 139 | 140 | 141 |
|---|---|---|---|---|---|---|---|---|---|
| A | 0.6340 | 0.9182 | 1.1924 | 1.0983 | 1.0868 | 0.7648 | 0.6656 | 1.2325 | 1.0404 |
| B | 1.9442 | 0.6660 | 1.8929 | 1.5843 | 1.6248 | 2.2051 | 1.8932 | 1.7917 | 2.5579 |
| C | 0.1197 | 0.2861 | 0.2393 | 0.1046 | 0.1535 | 0.1907 | 0.1259 | 0.0326 | 0.0499 |
| **Ability Categories** | | | | | | | | | |
| > 3. | - 0.26 | + 0.16 | + 0.62 | + 0.52 | + 0.54 | + 0.75 | + 0.46 | + 0.52 | - 3.03 |
| 2.8 | + 0.01 | - 0.74 | - 0.01 | - 0.18 | - 7.49 | + 1.72 | + 1.22 | + 0.49 | - 0.65 |
| 2.4 | + 0.08 | - 0.04 | + 0.19 | - 0.55 | + 3.64 | - 0.46 | - 1.47 | + 1.72 | - 0.00 |
| 2.0 | - 0.01 | - 0.01 | + 0.01 | + 0.00 | - 0.19 | + 0.04 | - 0.03 | - 0.67 | + 1.64 |
| 1.6 | - 0.47 | + 0.12 | - 0.27 | - 0.85 | - 0.02 | + 0.77 | + 0.02 | - 0.09 | + 1.66 |
| 1.2 | + 0.14 | + 0.60 | - 1.36 | + 0.67 | + 0.01 | - 0.00 | + 0.22 | - 0.12 | + 0.15 |
| .8 | + 3.63 | + 0.00 | + 0.93 | + 1.51 | - 0.18 | - 0.05 | - 0.36 | - 0.02 | - 0.97 |
| .4 | - 2.66 | - 0.02 | + 0.78 | - 0.45 | + 0.31 | + 0.00 | + 0.44 | - 0.06 | - 8.23 |
| .0 | - 0.05 | - 1.74 | + 0.37 | - 0.16 | - 0.34 | + 0.50 | - 0.02 | + 4.73 | + 0.00 |
| -.4 | - 0.26 | + 0.35 | - 2.26 | - 1.72 | + 1.99 | + 2.84 | + 0.25 | - 0.25 | - 0.43 |
| -.8 | + 0.39 | + 0.01 | - 0.09 | - 0.29 | - 1.08 | - 1.27 | - 0.07 | + 2.37 | + 0.10 |
| -1.2 | + 0.17 | + 3.67 | + 0.47 | + 0.01 | - 0.13 | - 0.07 | - 3.12 | - 2.47 | + 5.25 |
| -1.6 | - 0.22 | + 0.09 | + 3.57 | + 0.18 | + 1.42 | + 0.40 | + 0.30 | - 2.78 | + 7.55 |
| -2.0 | + 0.00 | - 0.40 | + 0.47 | + 1.91 | + 0.06 | - 1.13 | + 0.22 | - 0.05 | + 0.13 |
| -2.4 | - 0.05 | - 0.62 | + 0.01 | + 2.40 | + 0.85 | + 0.11 | + 0.07 | + 6.01 | - 0.58 |
| -2.8 | + 1.14 | + 1.73 | - 1.37 | + 23.05 | + 0.57 | + 0.17 | + 1.01 | + 2.16 | + 0.95 |
| <-3. | + 0.29 | - 0.55 | + 0.01 | + 5.91 | + 0.38 | - 2.16 | + 3.88 | - 0.40 | - 0.53 |
| TOTAL CHI | 9.83 | 10.76 | 12.80 | 40.37 | 19.21 | 12.44 | 13.16 | 24.82 | 31.86 |

[1]Items appear in the following order:
PSAT/NMSQT Form 1, items 1-50;
PSAT/NMSQT Form 2, items 51-100;
SAT First Old Form, items 101-120;
SAT Second Old Form, items 121-141.

[2]Asterisks next to item numbers indicate items with questionable fit to the three-parameter logistic model.

Figure 6: PSAT/NMSQT Form 1 Item Ability Regression Plots

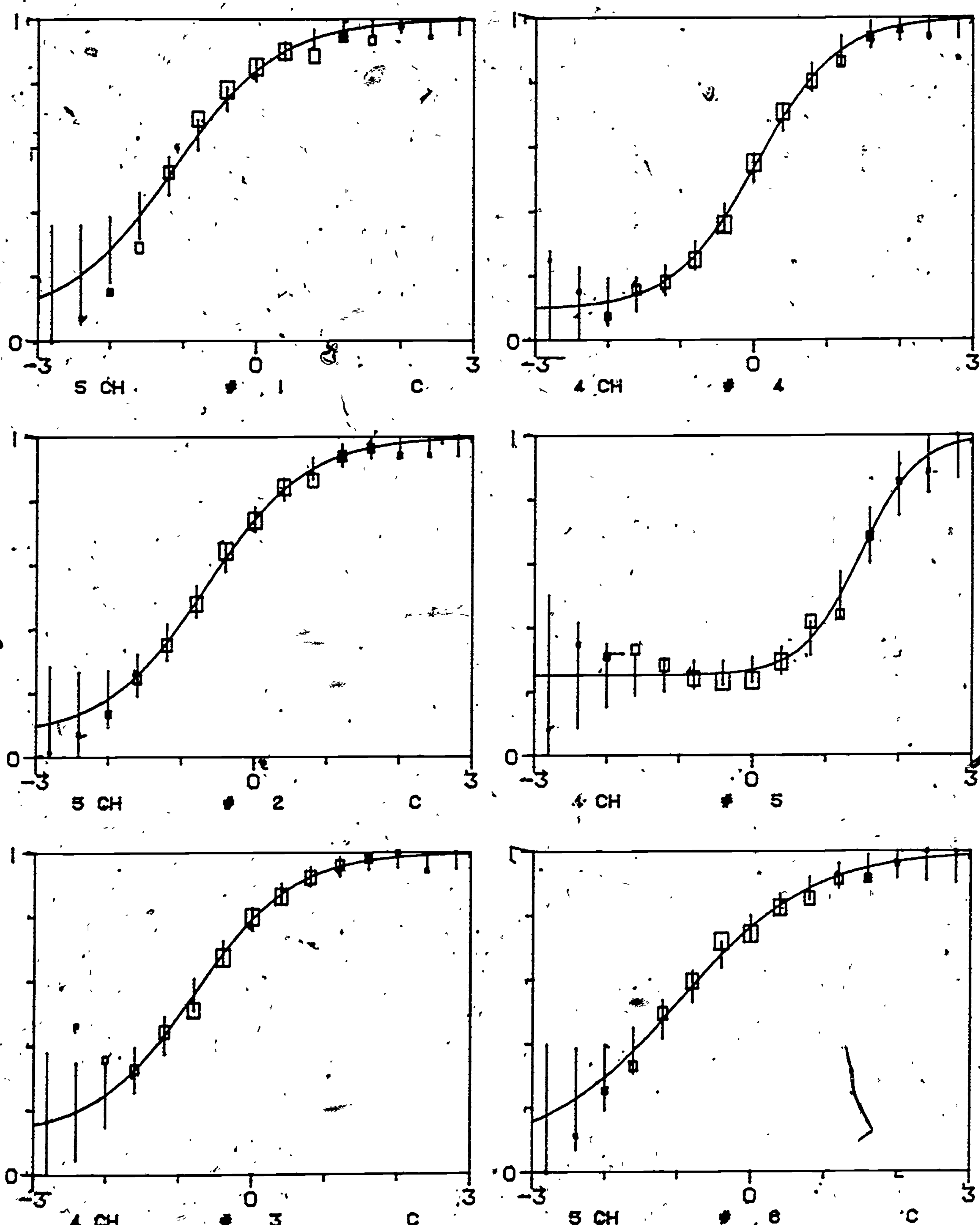


Figure 7: PSAT/NMSQT Form 1 Item Ability Regression Plots

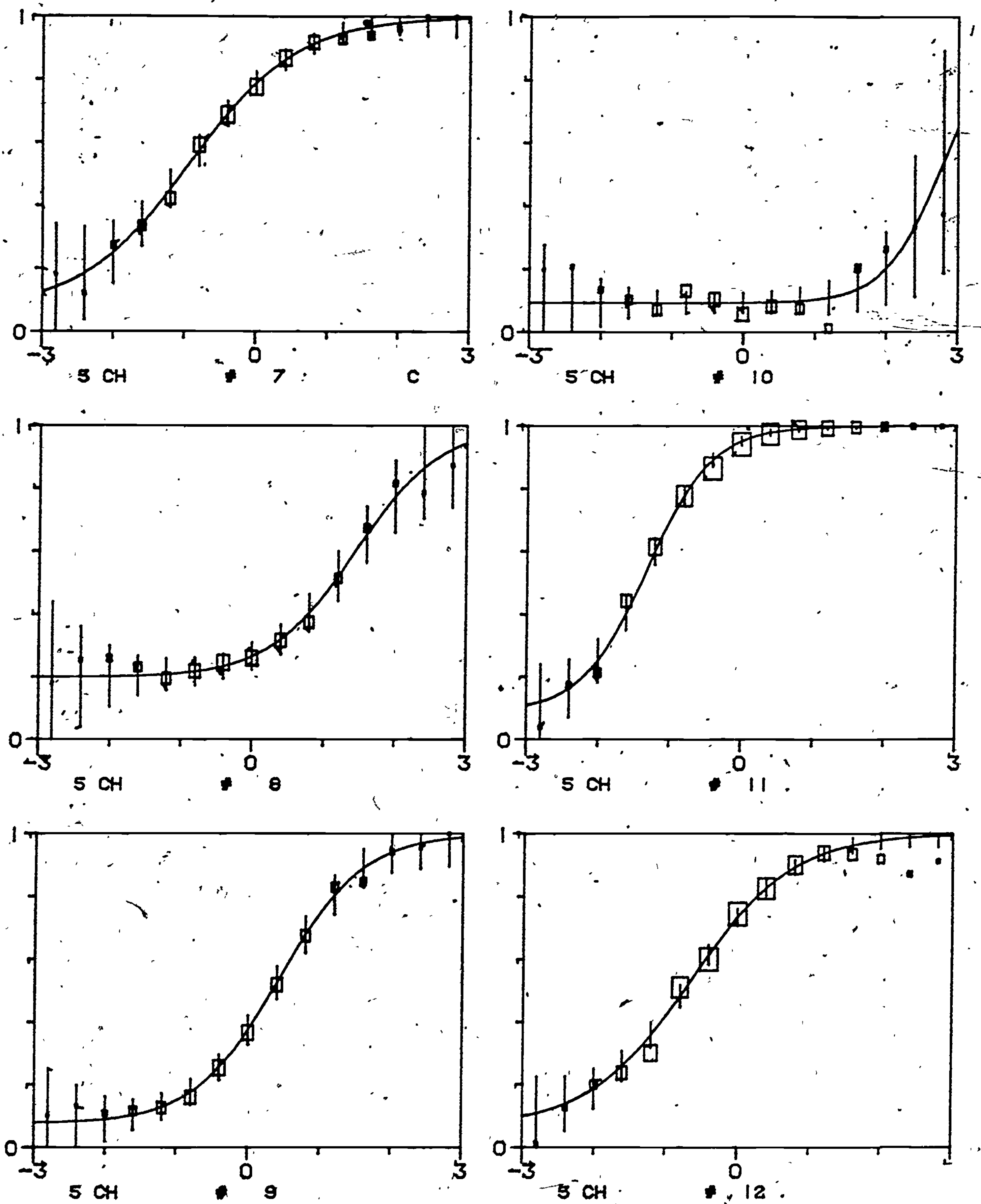


Figure 8: PSAT/NMSQT Form 1 Item Ability Regression Plots.

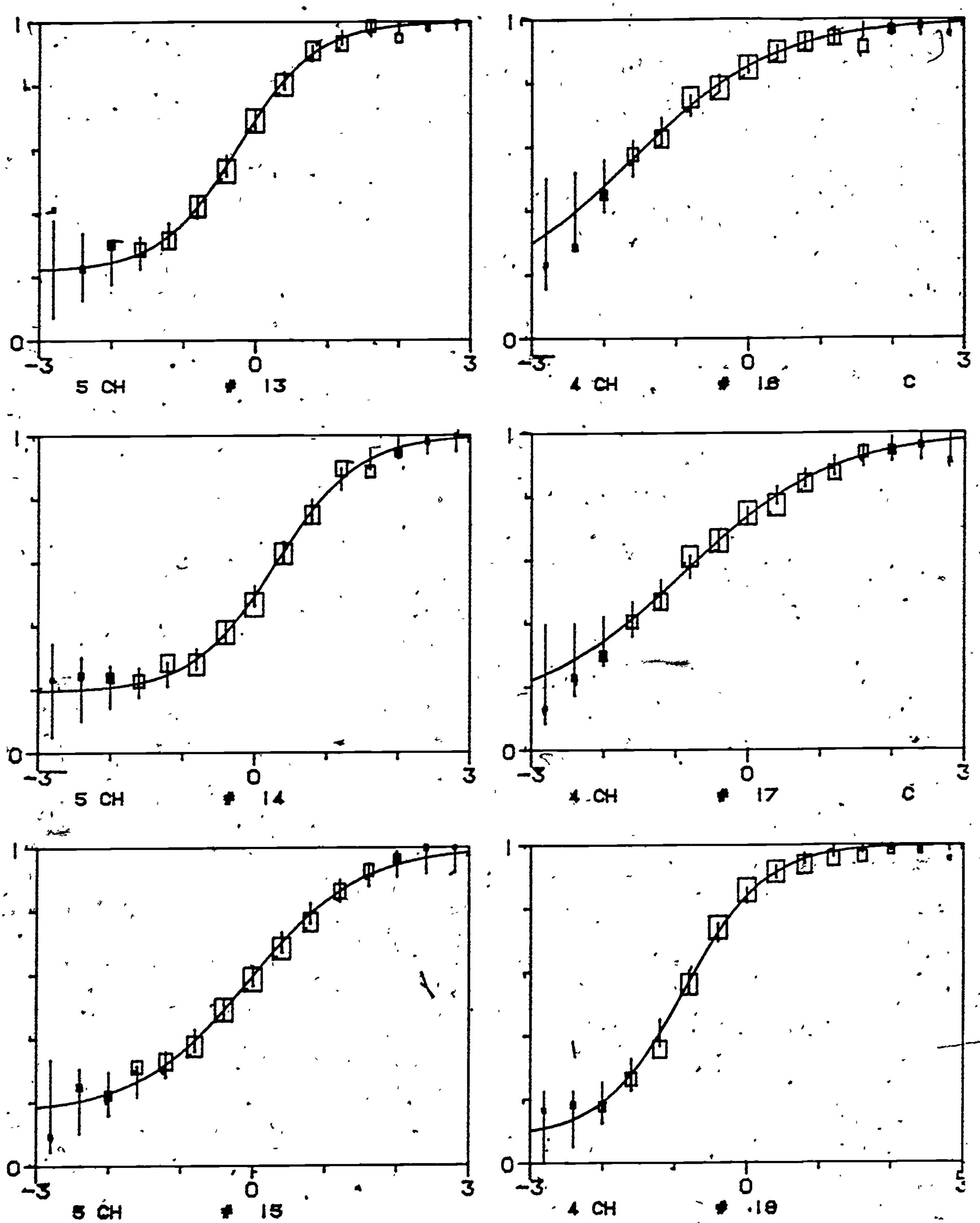


Figure 9: PSAT/NMSQT Form 1 Item Ability Regression Plots

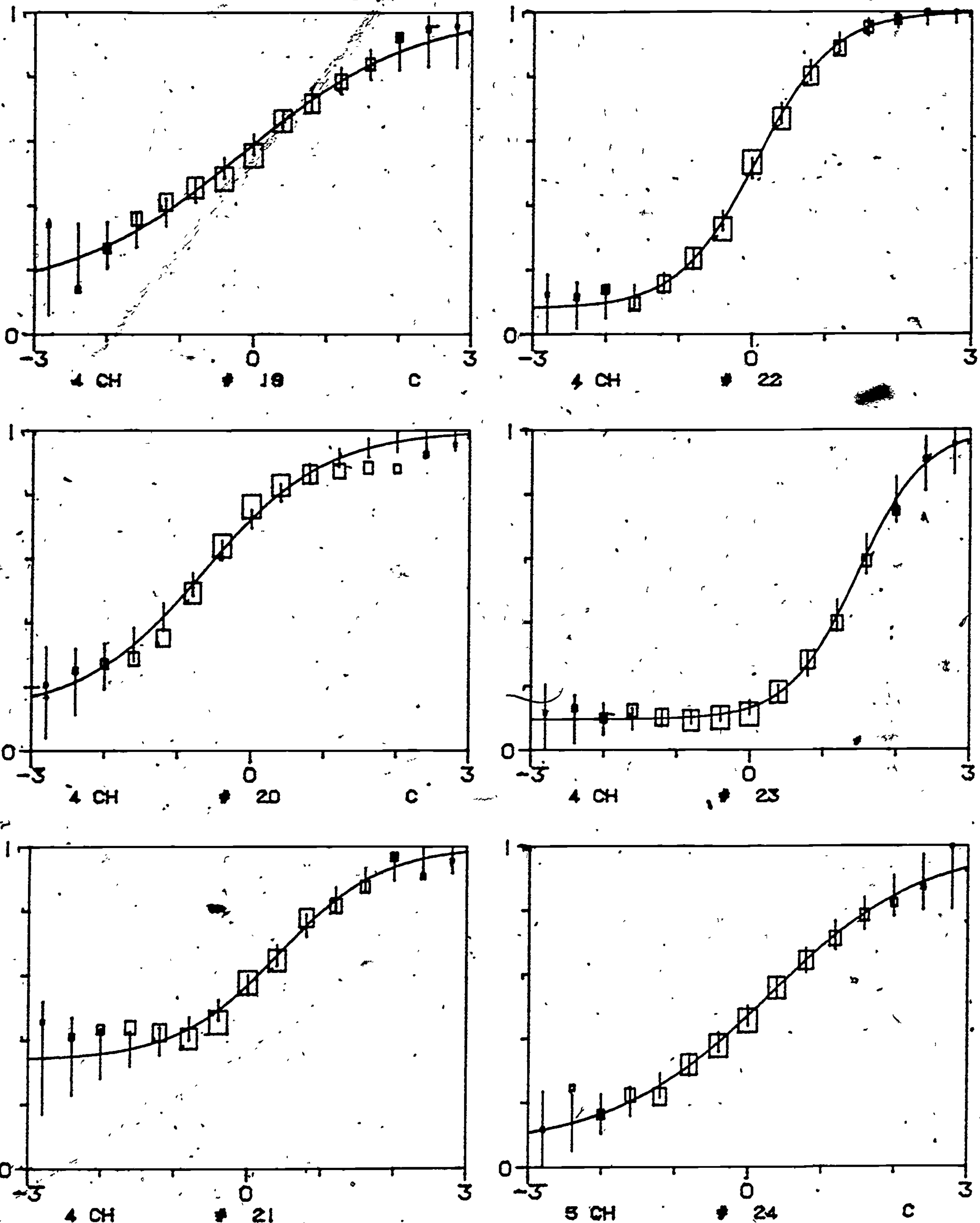


Figure 10: PSAT/NMSQT Form 1 Item Ability Regression Plots

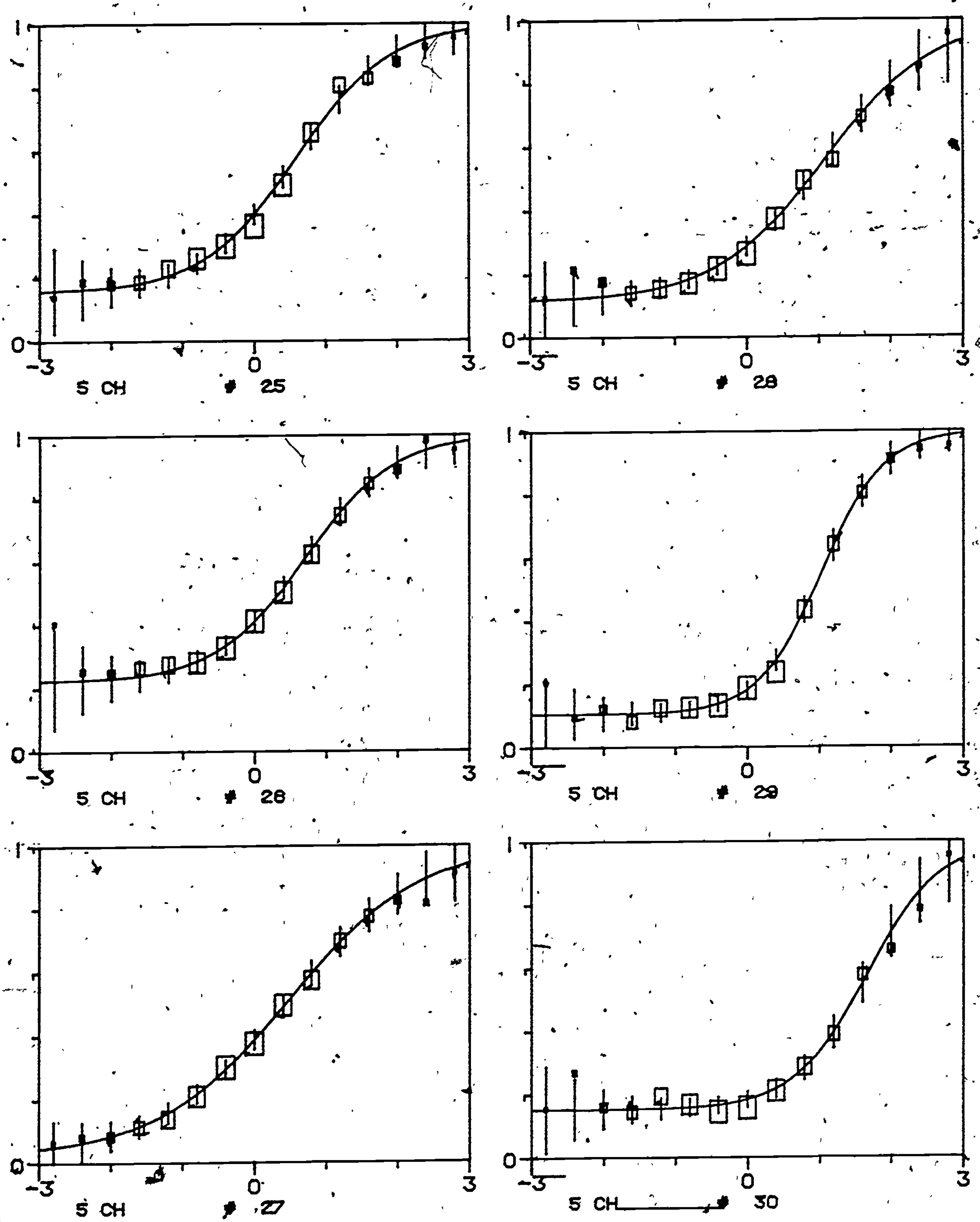


Figure 11: PSAT/NMSQT Form 1 Item Ability Regression Plots

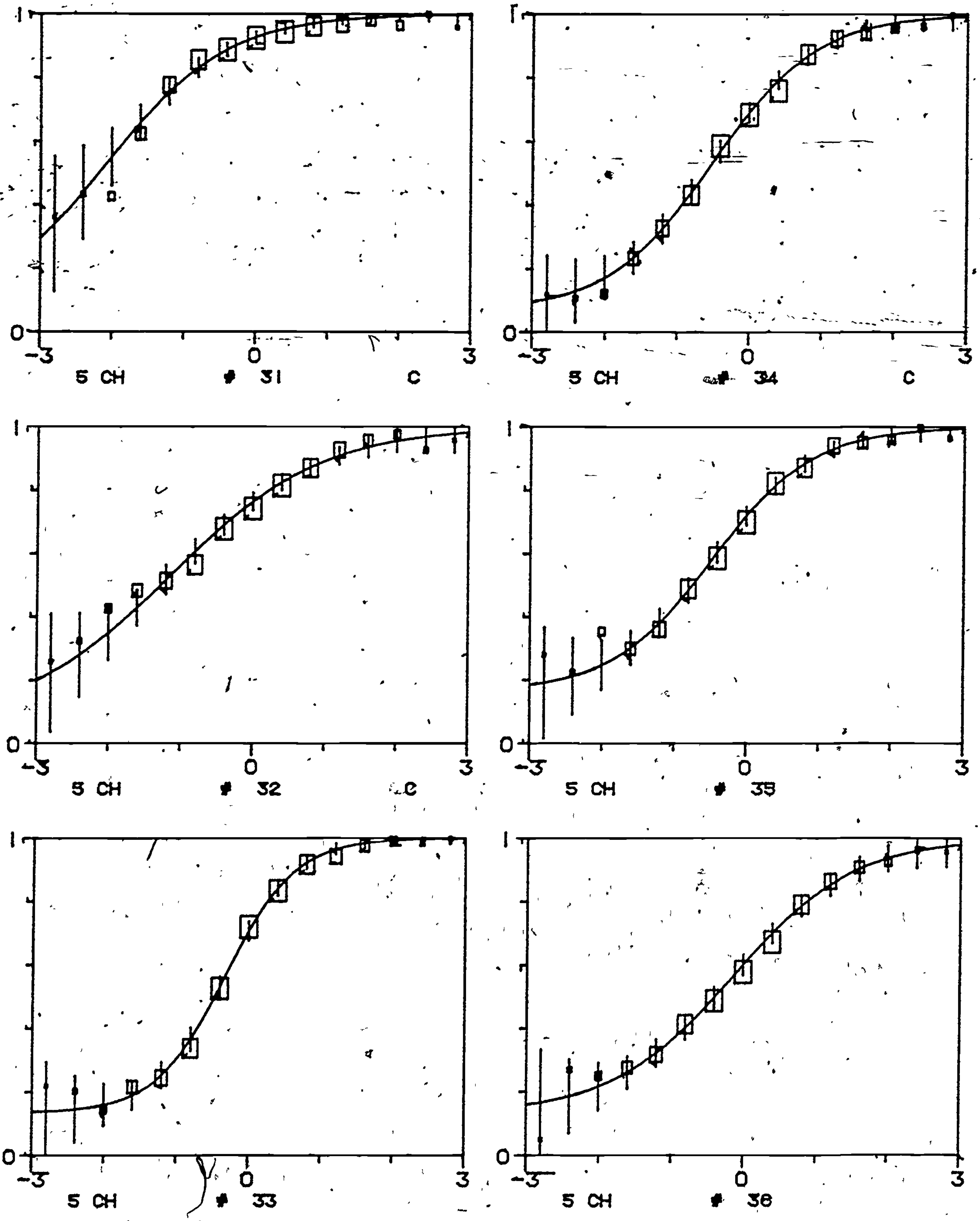


Figure 12: PSAT/NMSQT Form 1 Item Ability Regression Plots

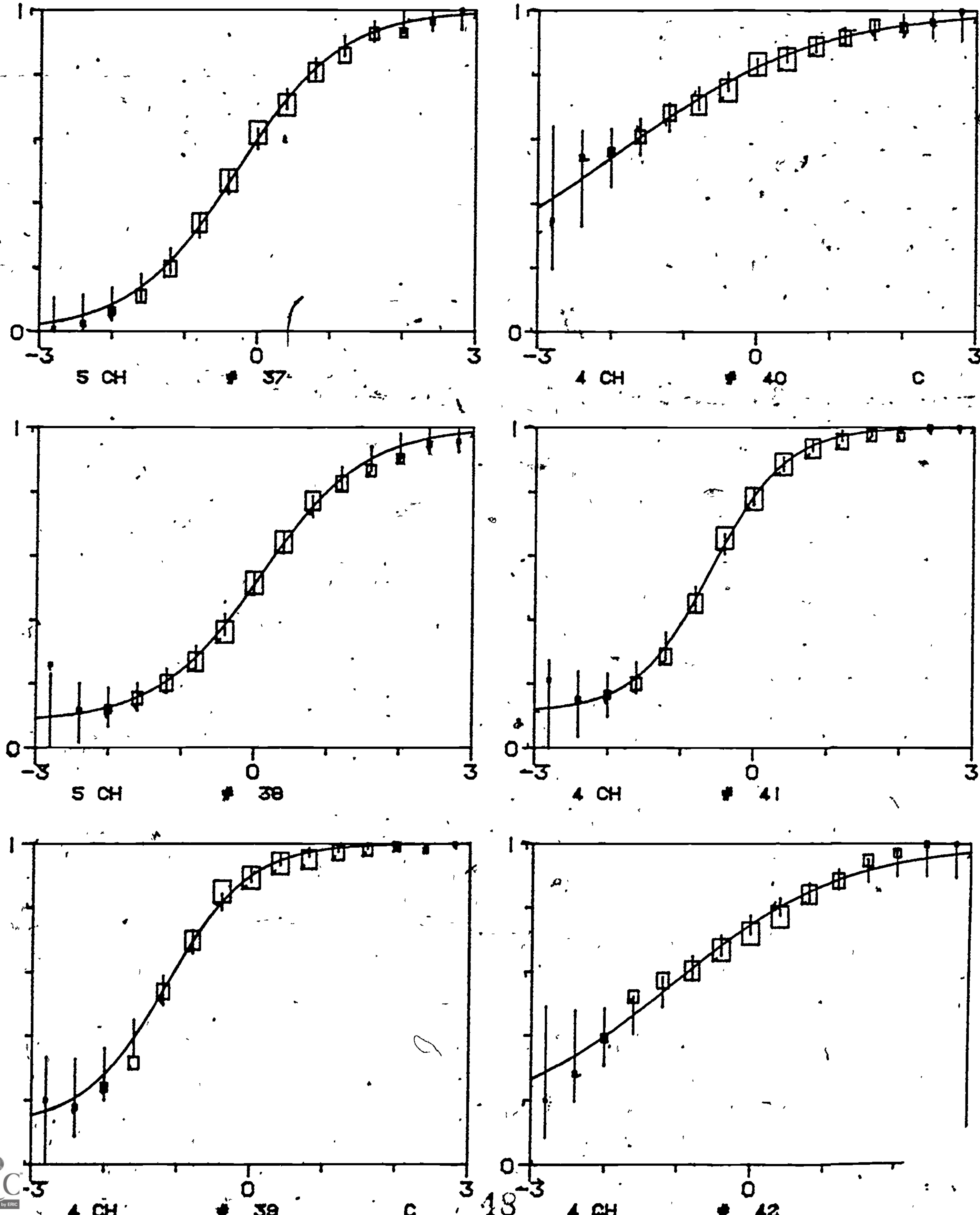


Figure 13: PSAT/NMSQT Form 1 Item Ability Regression Plots

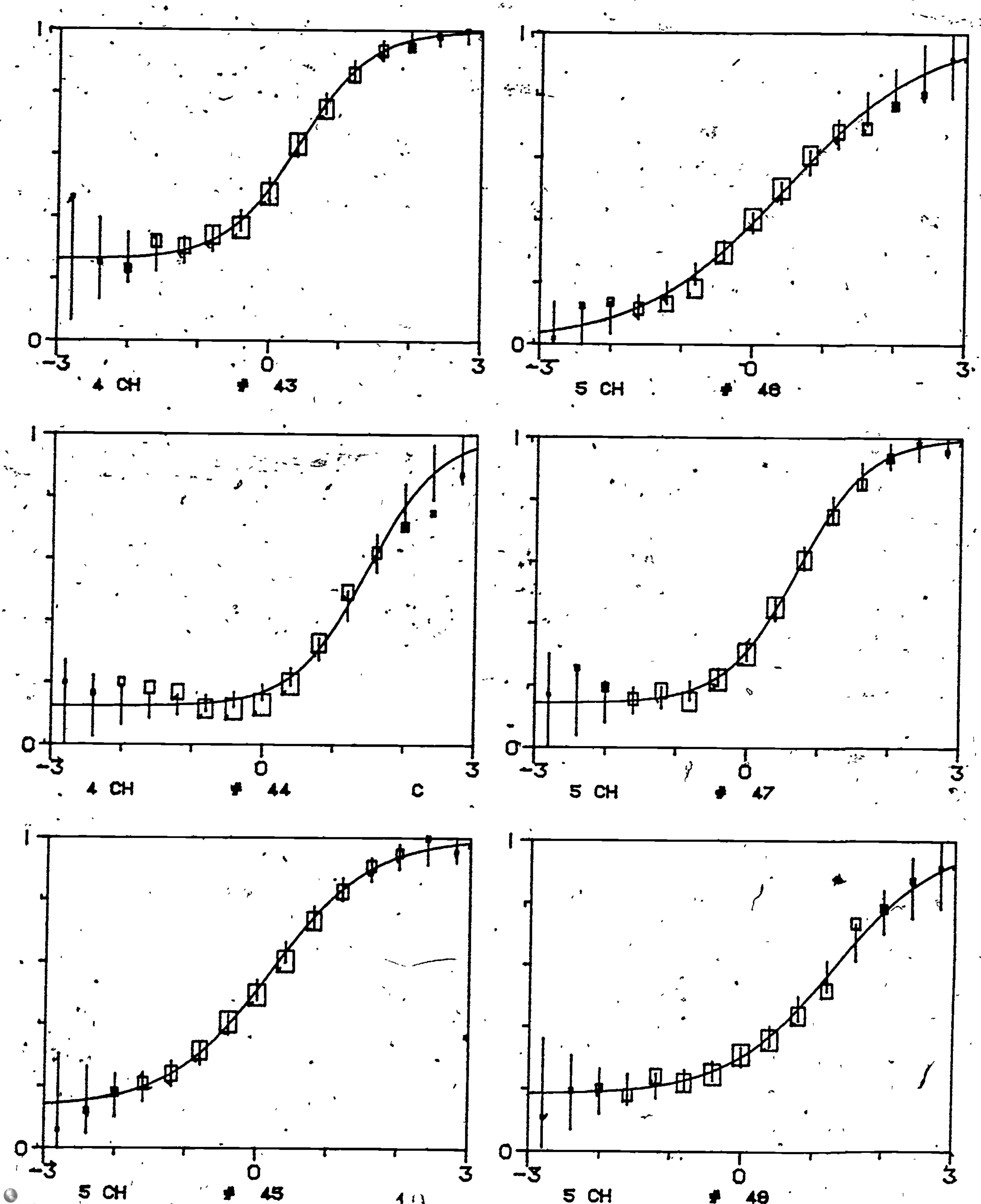
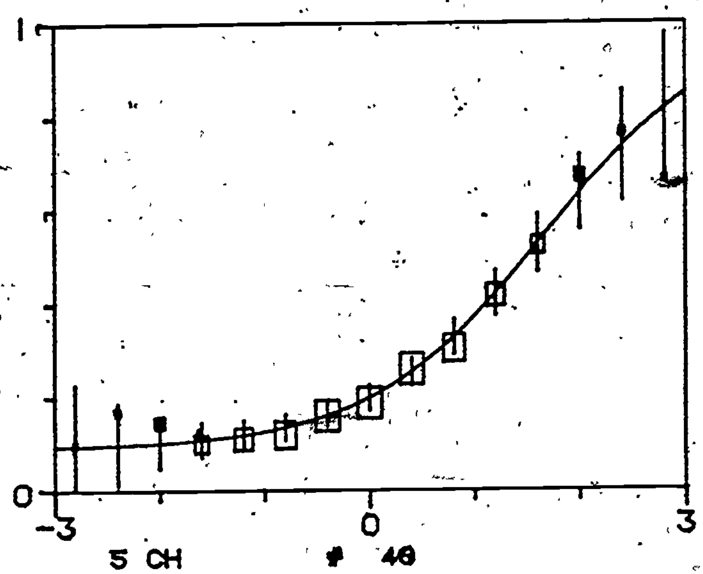


Figure 14. PSAT/NMSQT Form 1 Item Ability Regression Plots



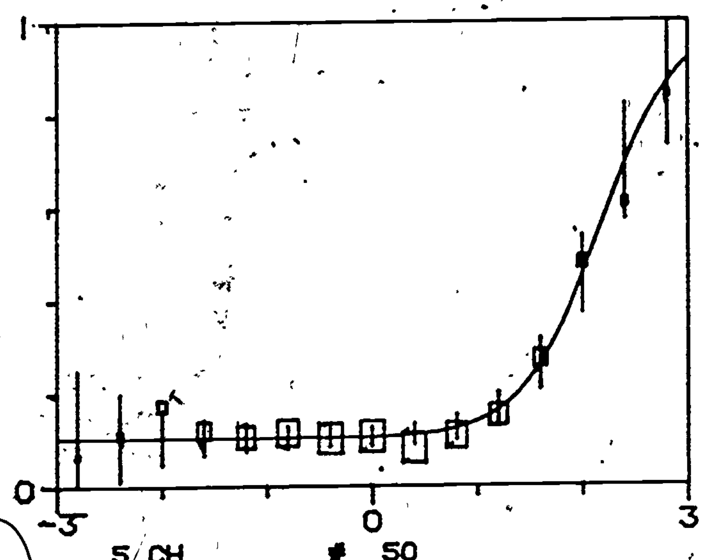



Figure 15: PSAT/NMSQT Form 2 Item Ability Regression Plots

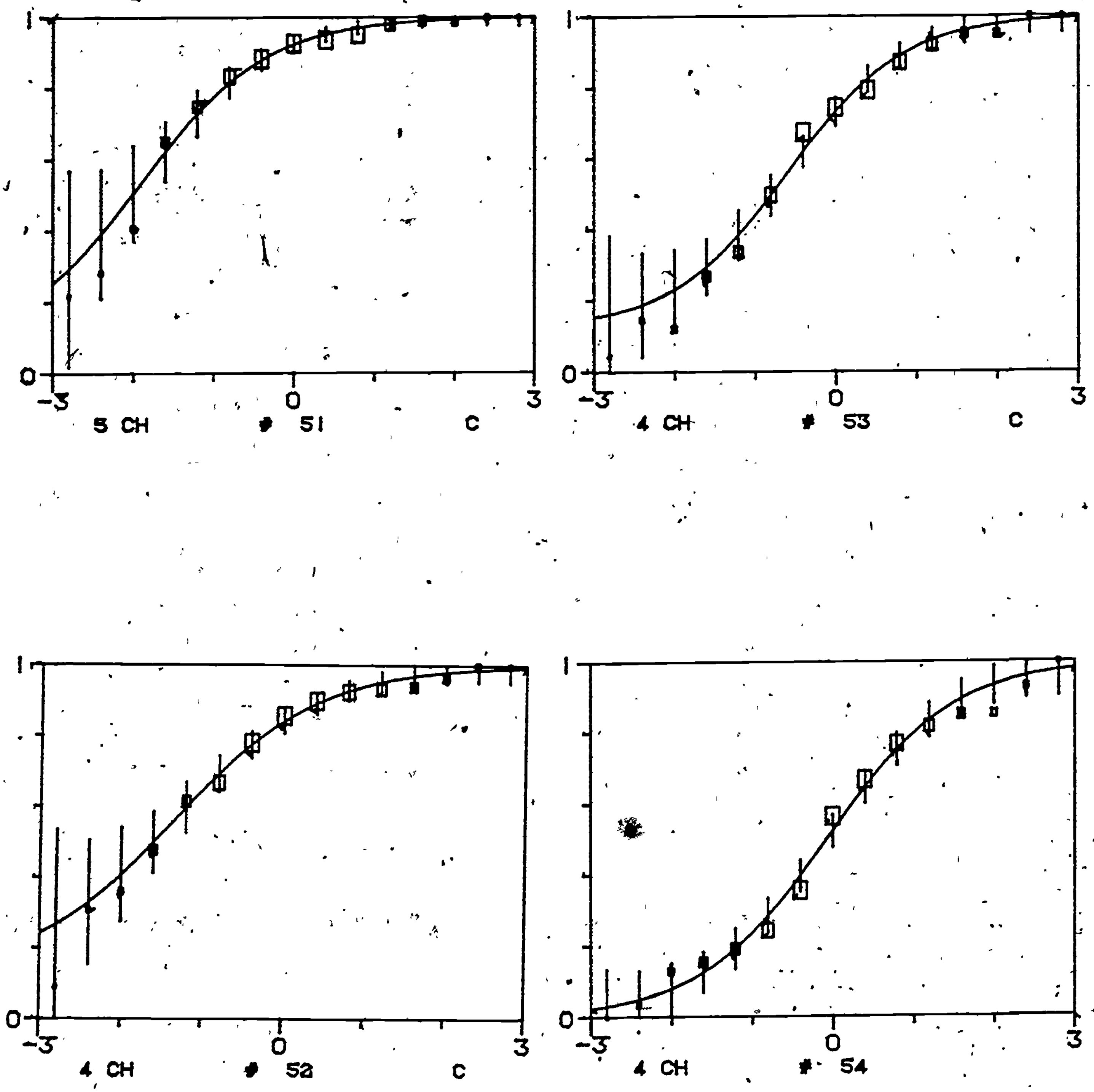


Figure 16: PSAT/NMSQT Form 2 Item Ability Regression Plots

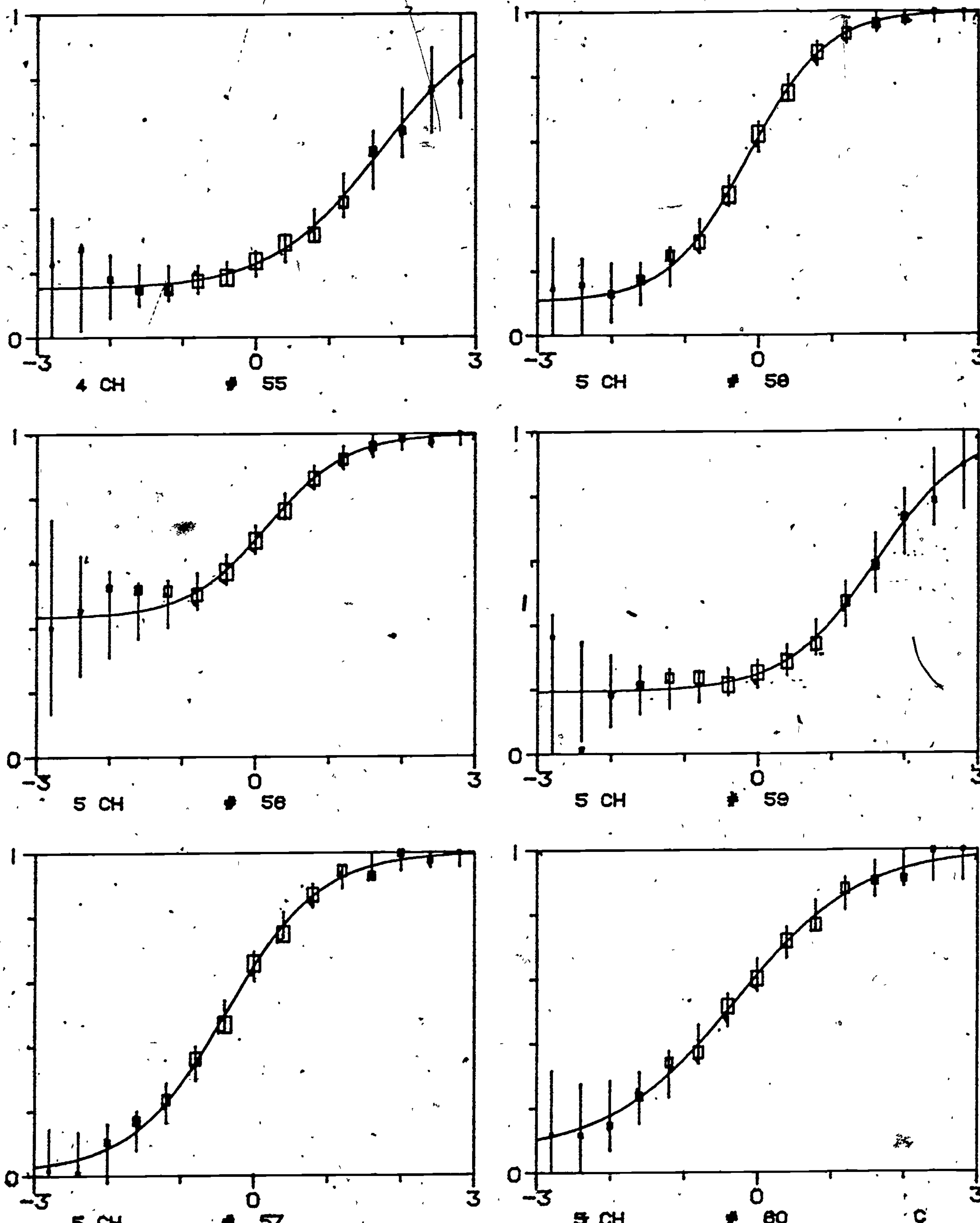


Figure 17: PSAT/NMSQT Form 2 Item Ability Regression Plots

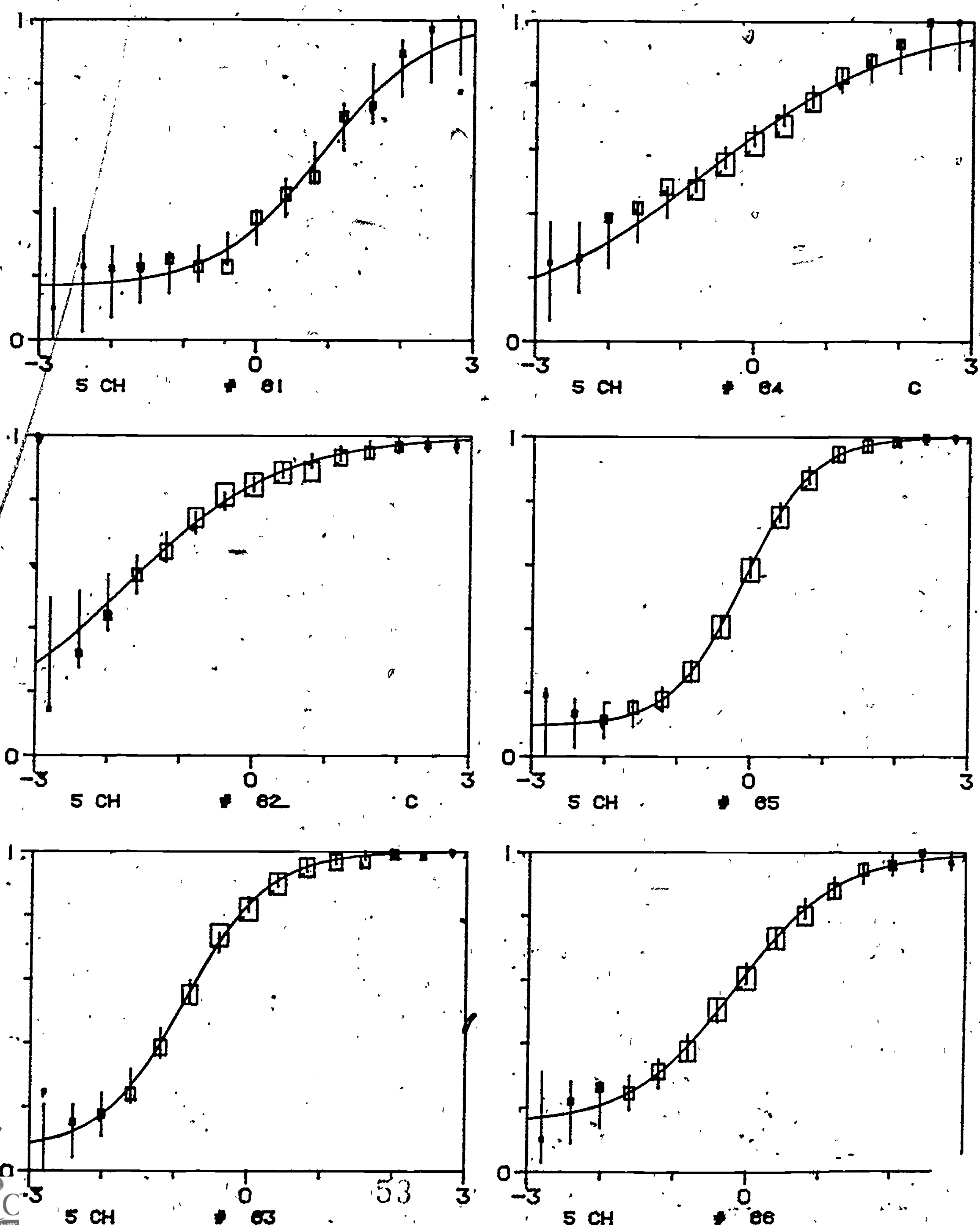


Figure 18: PSAT/NMSQT Form 2 Item Ability Regression Plots

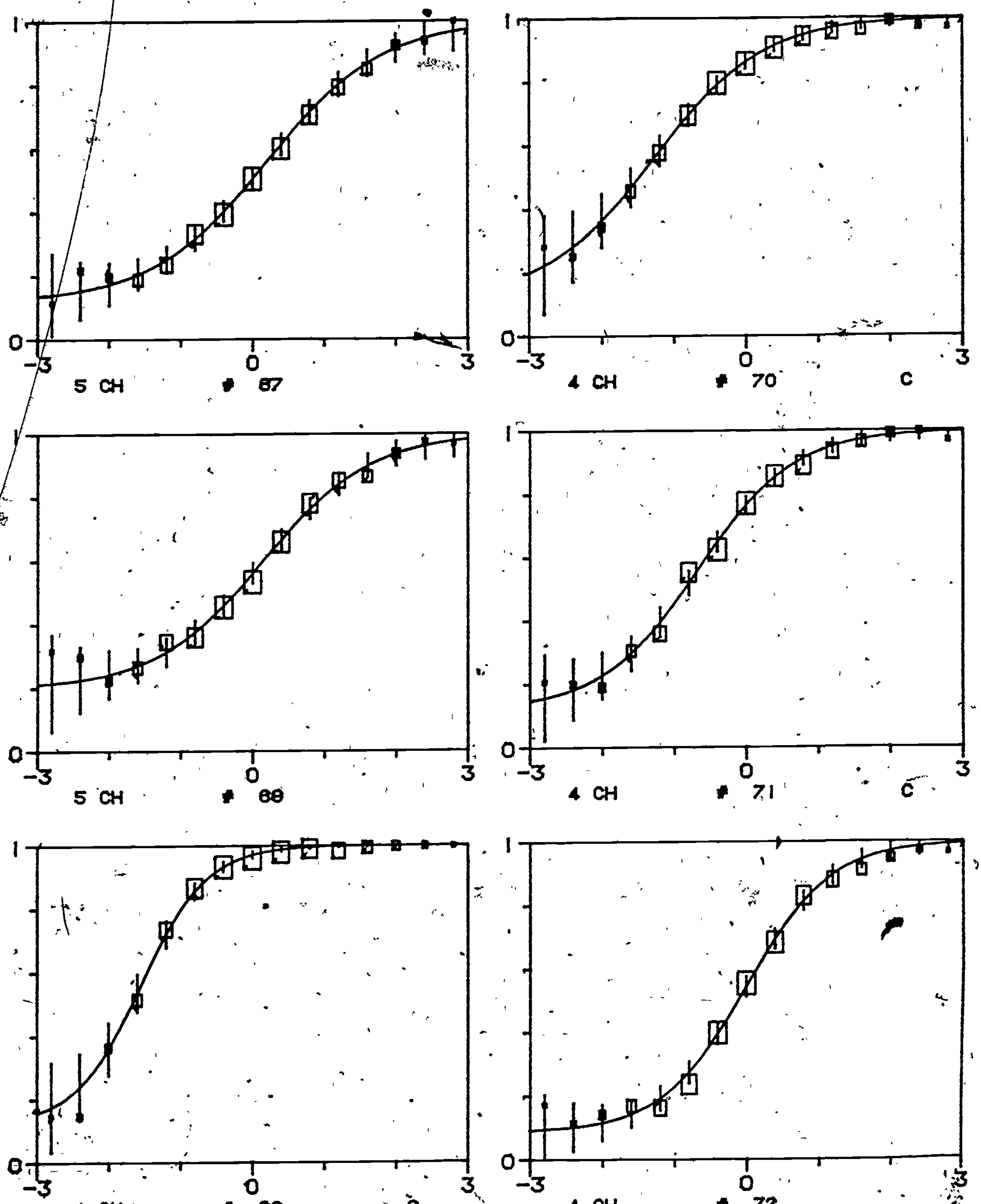


Figure 19: PSAT/NMSQT Form 2 Item Ability Regression Plots

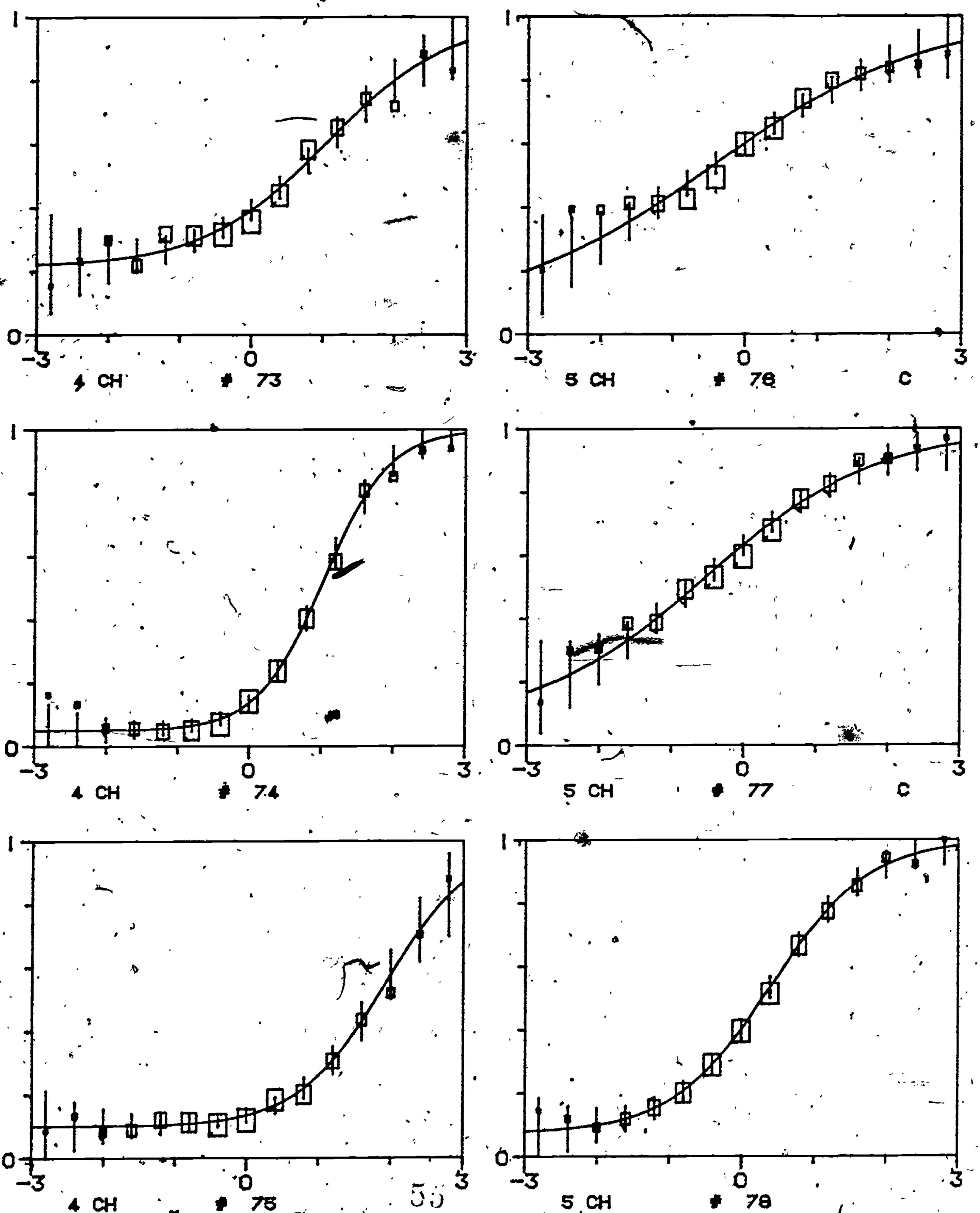


Figure 20: PSAT/NMSQT Form 2 Item Ability Regression Plots

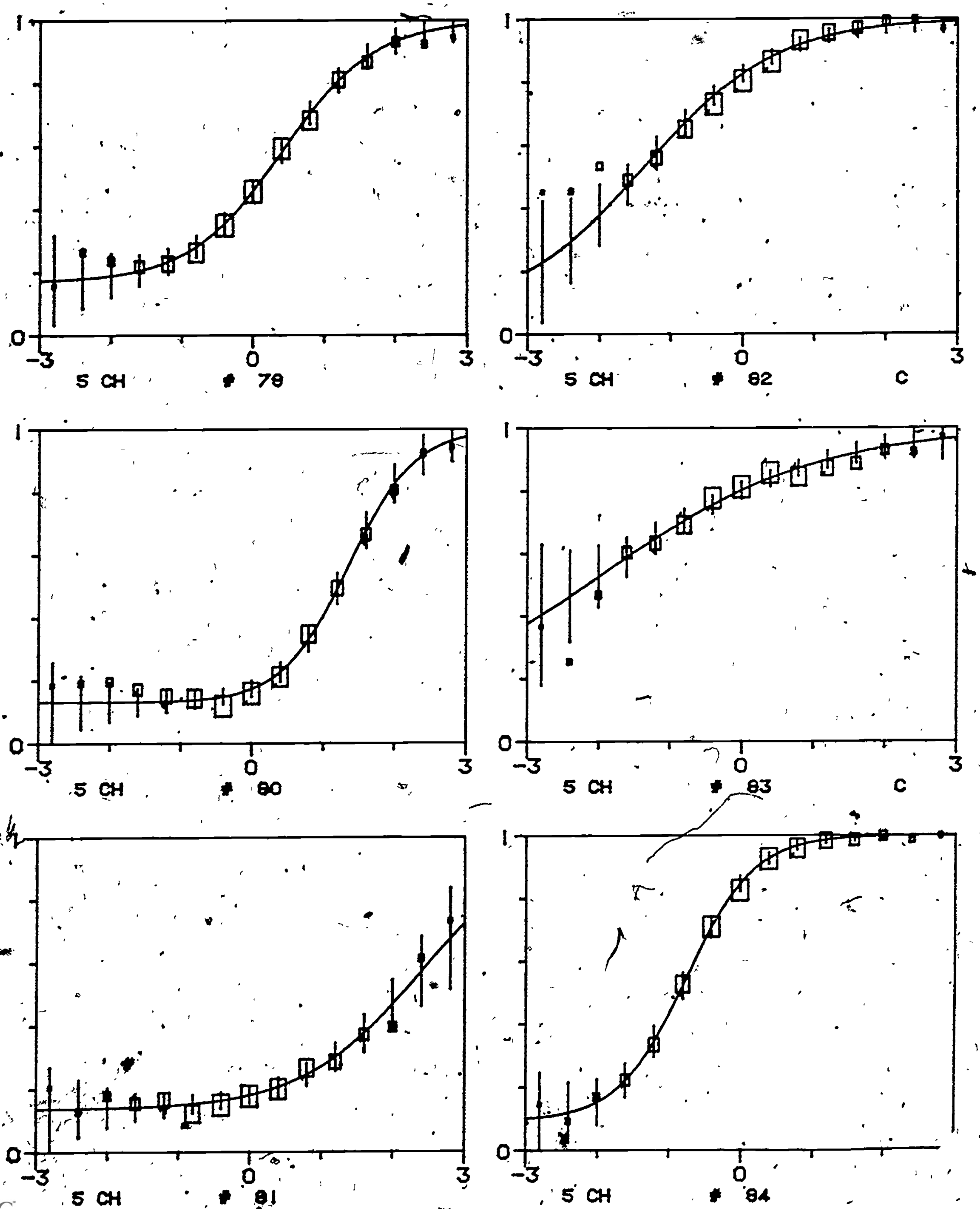


Figure 21: PSAT/NMSQT Form 2 Item Ability Regression Plots

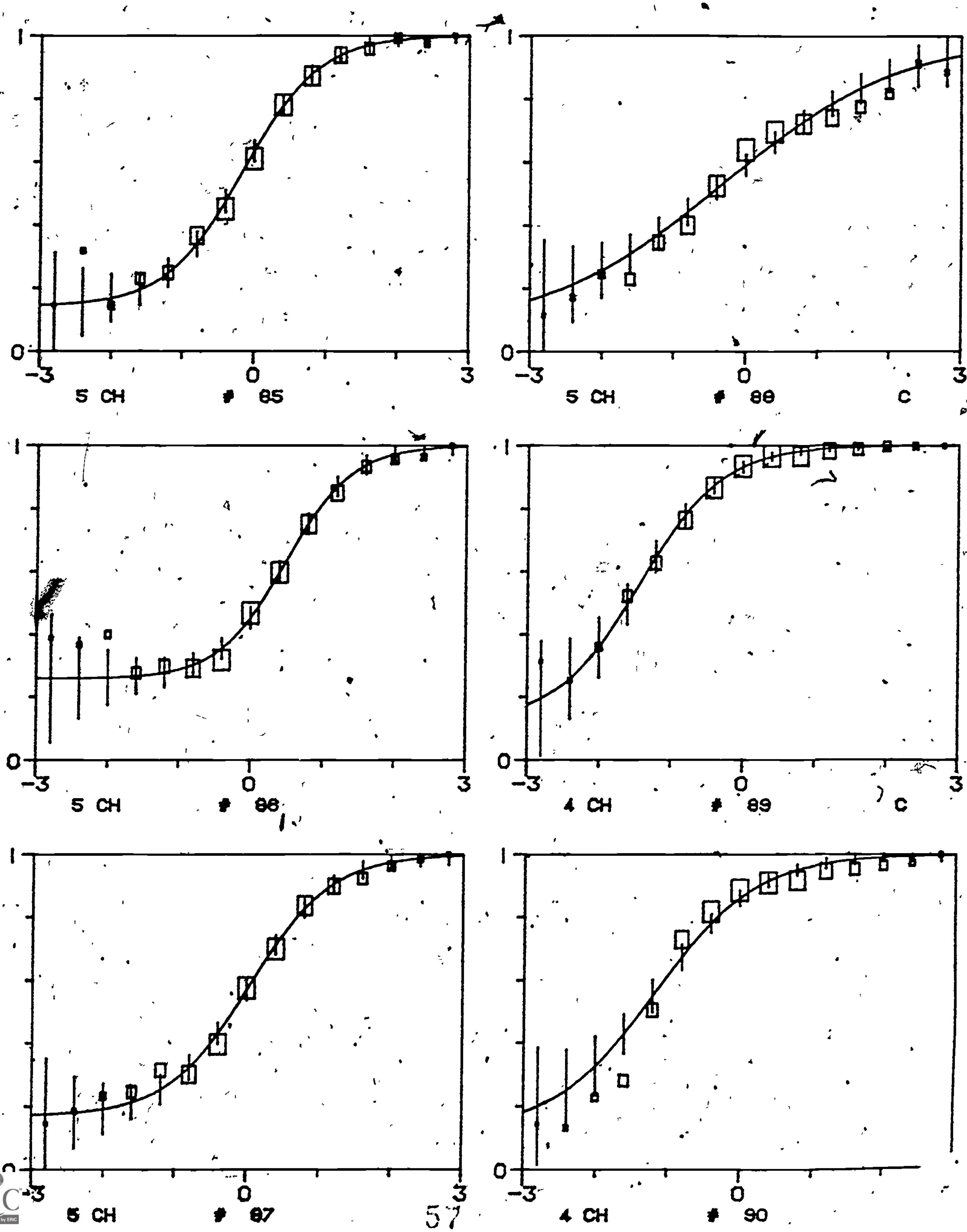


Figure 22: PSAT/NMSQT Form 2 Item Ability Regression Plots

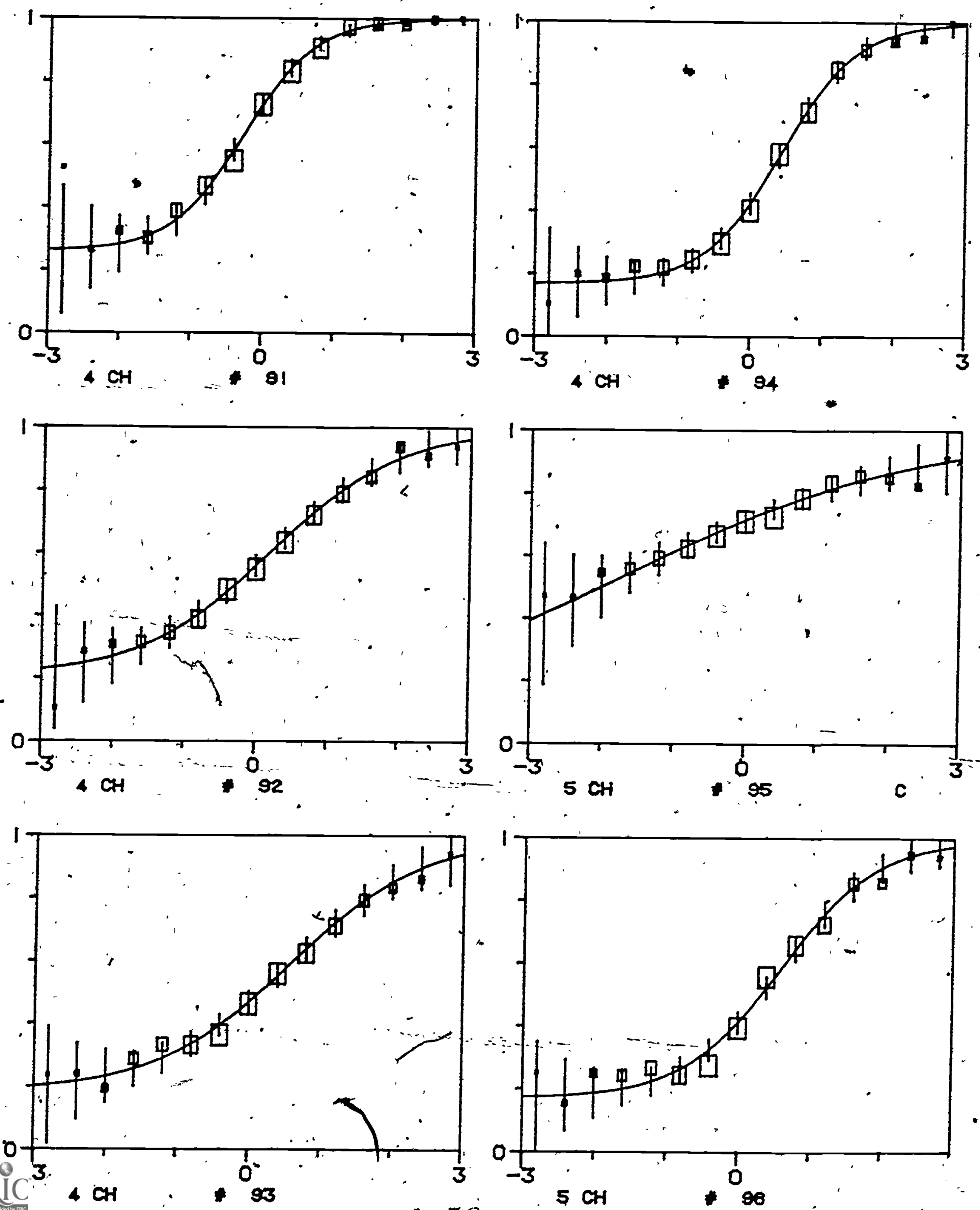


Figure 23: PSAT/NMSQT Form 2 Item Ability Regression Plots

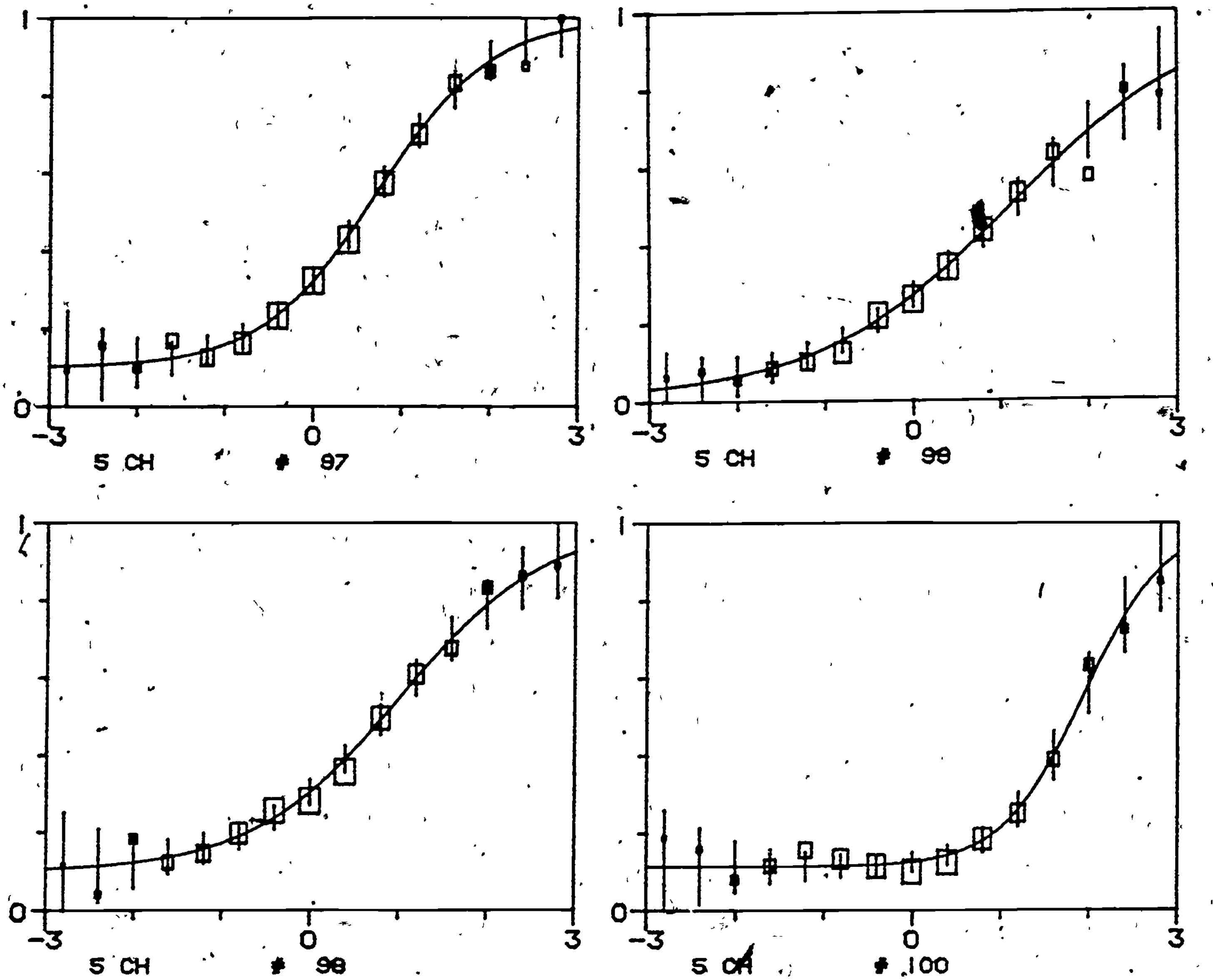
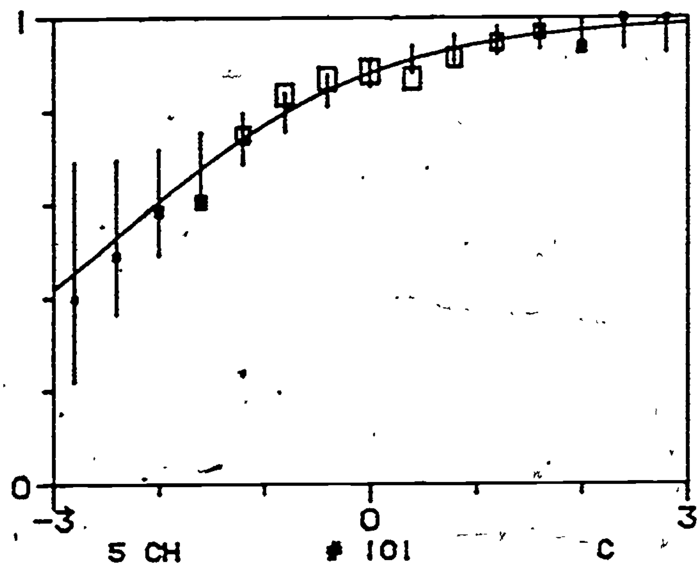


Figure 24: SAT First Old Form Item Ability Regression Plots



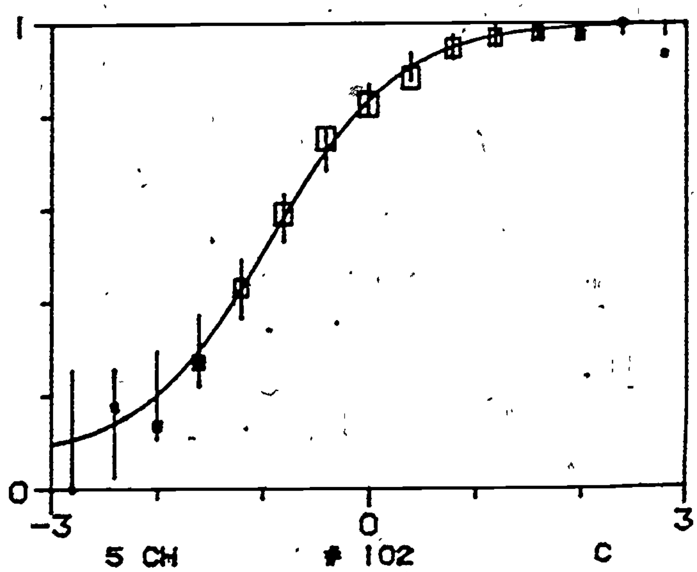



Figure 25: SAT First Old Form Item Ability Regression Plots

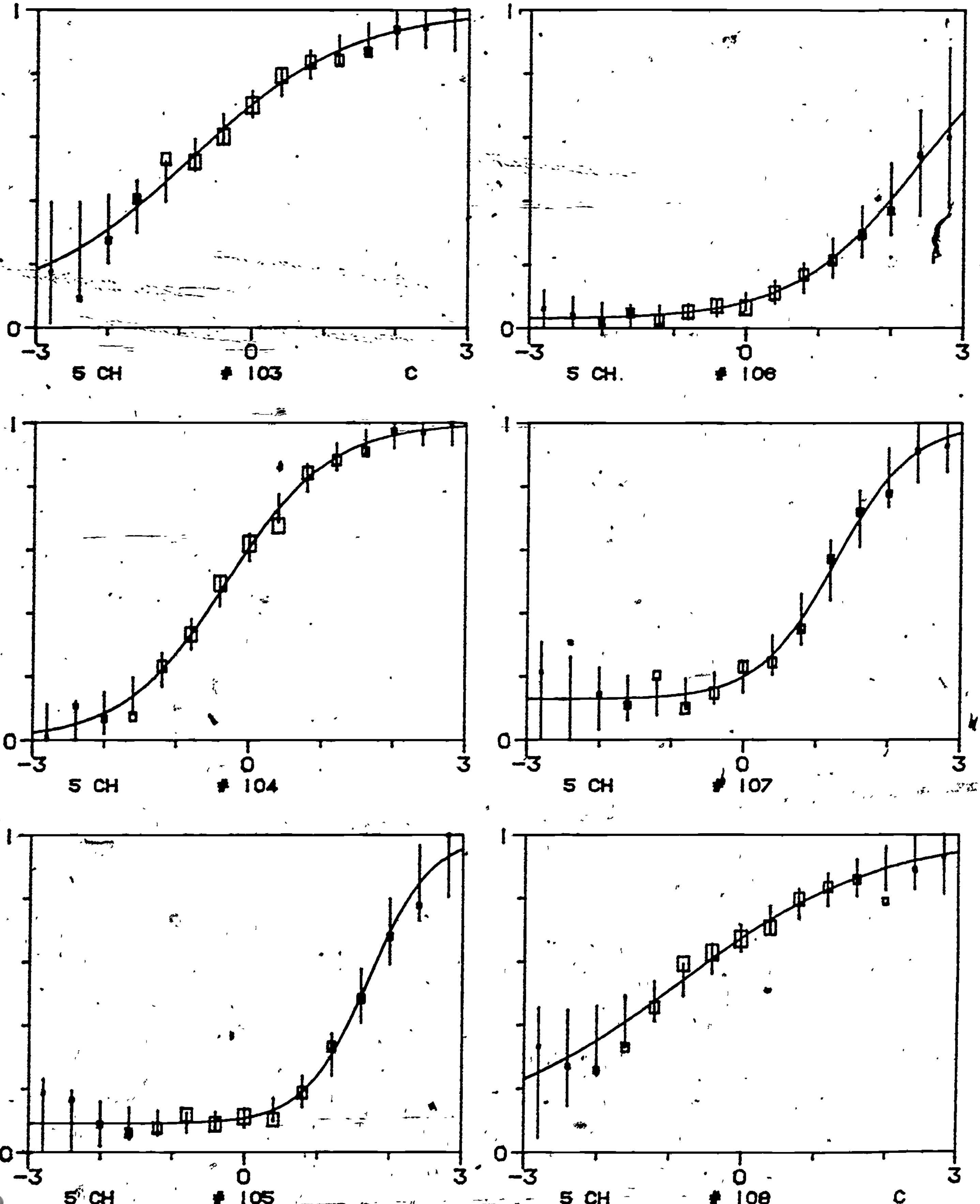


Figure 26: SAT First Old Form Item Ability Regression Plots

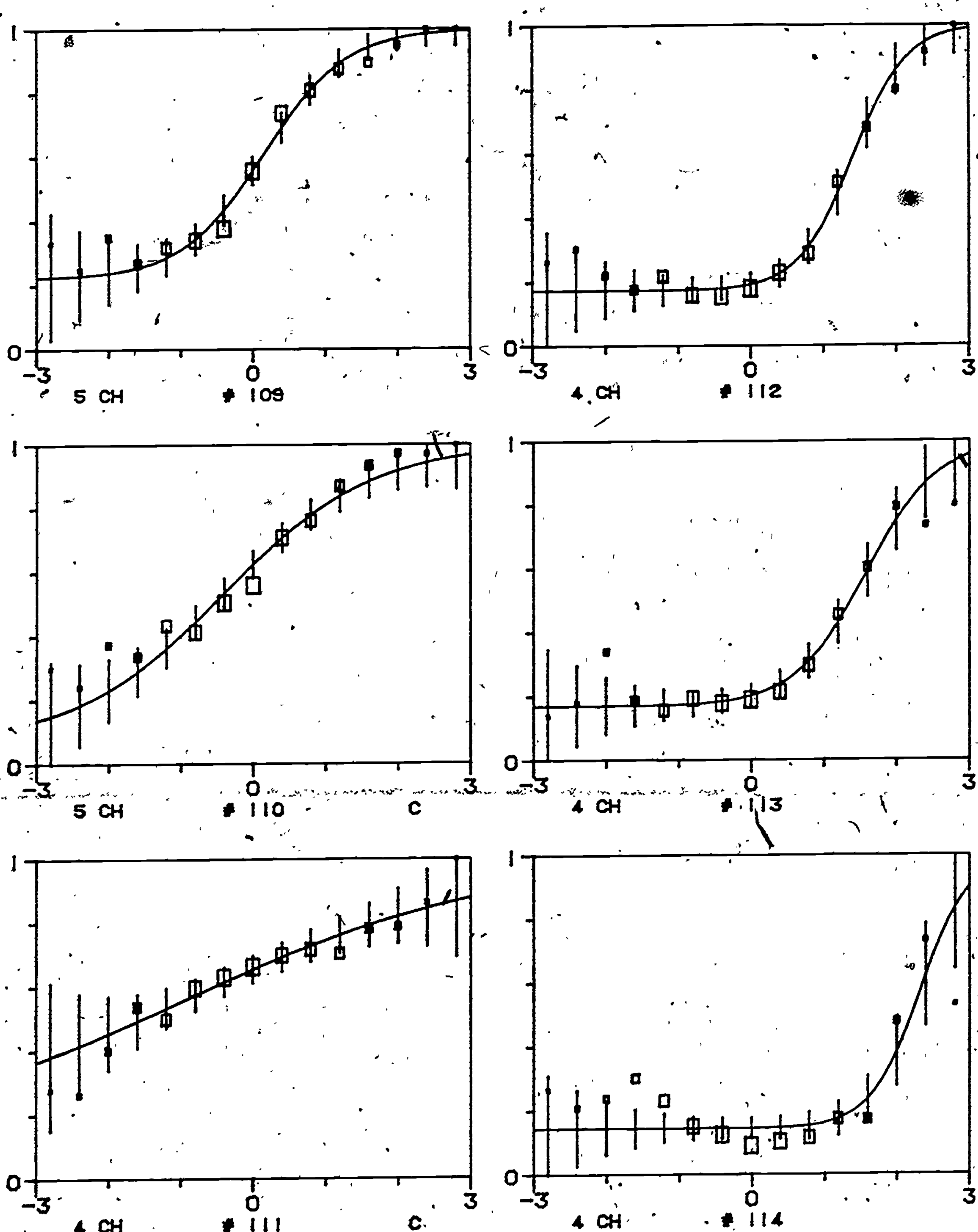


Figure 27: SAT First Old Form Item Ability Regression Plots

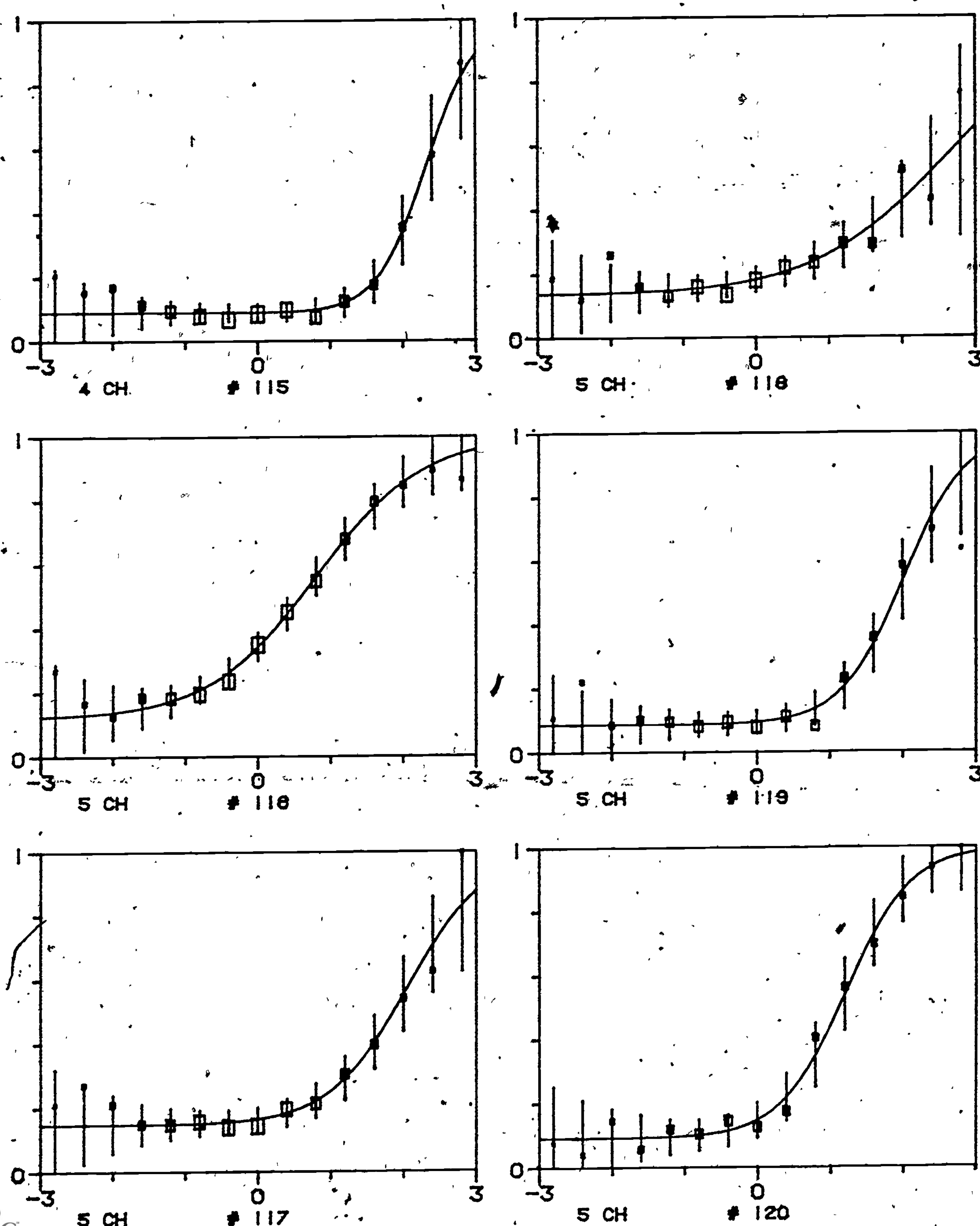


Figure 28: SAT Second Old Form Item Ability Regression Plots

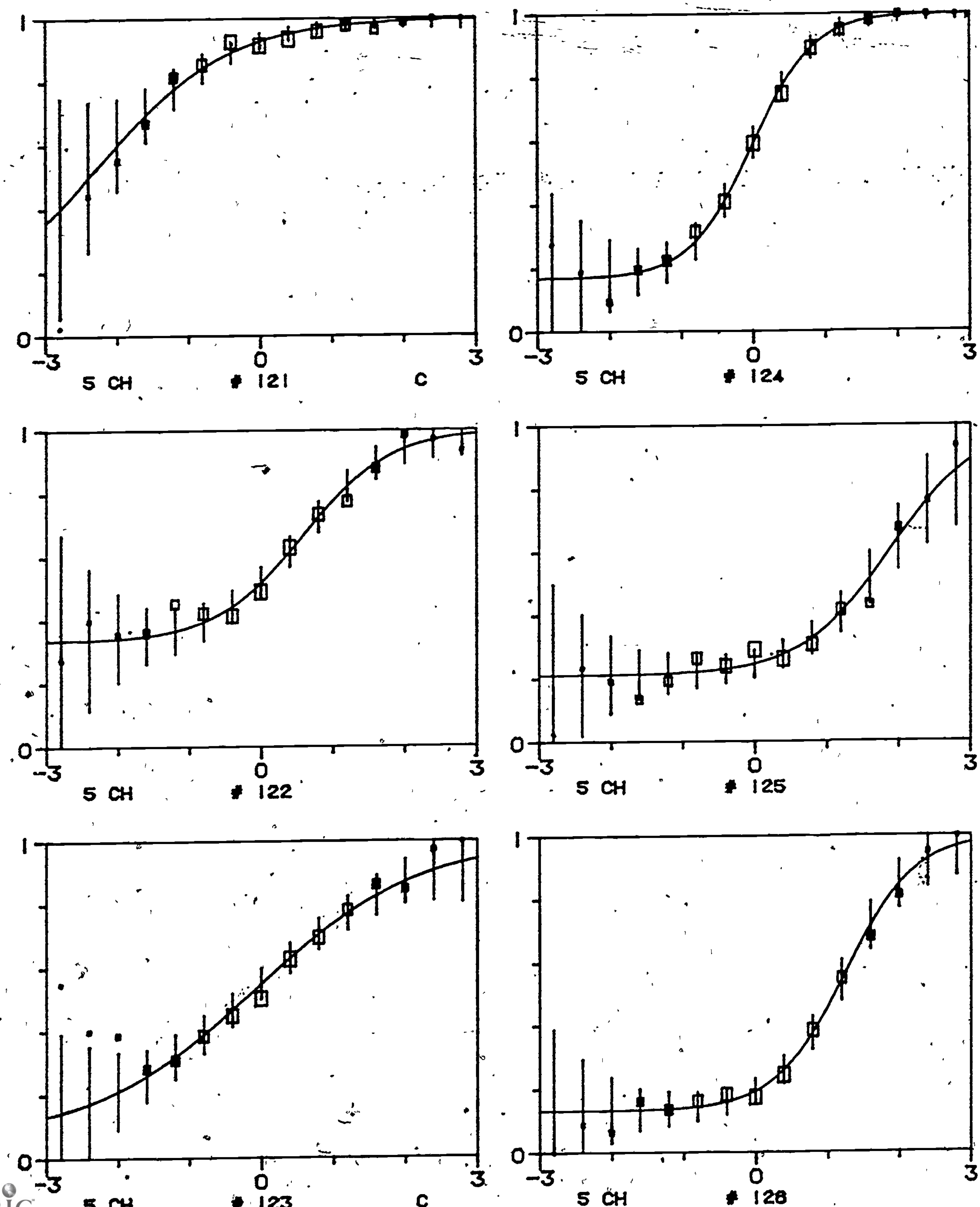


Figure 29: SAT Second Old Form Item Ability Regression Plots

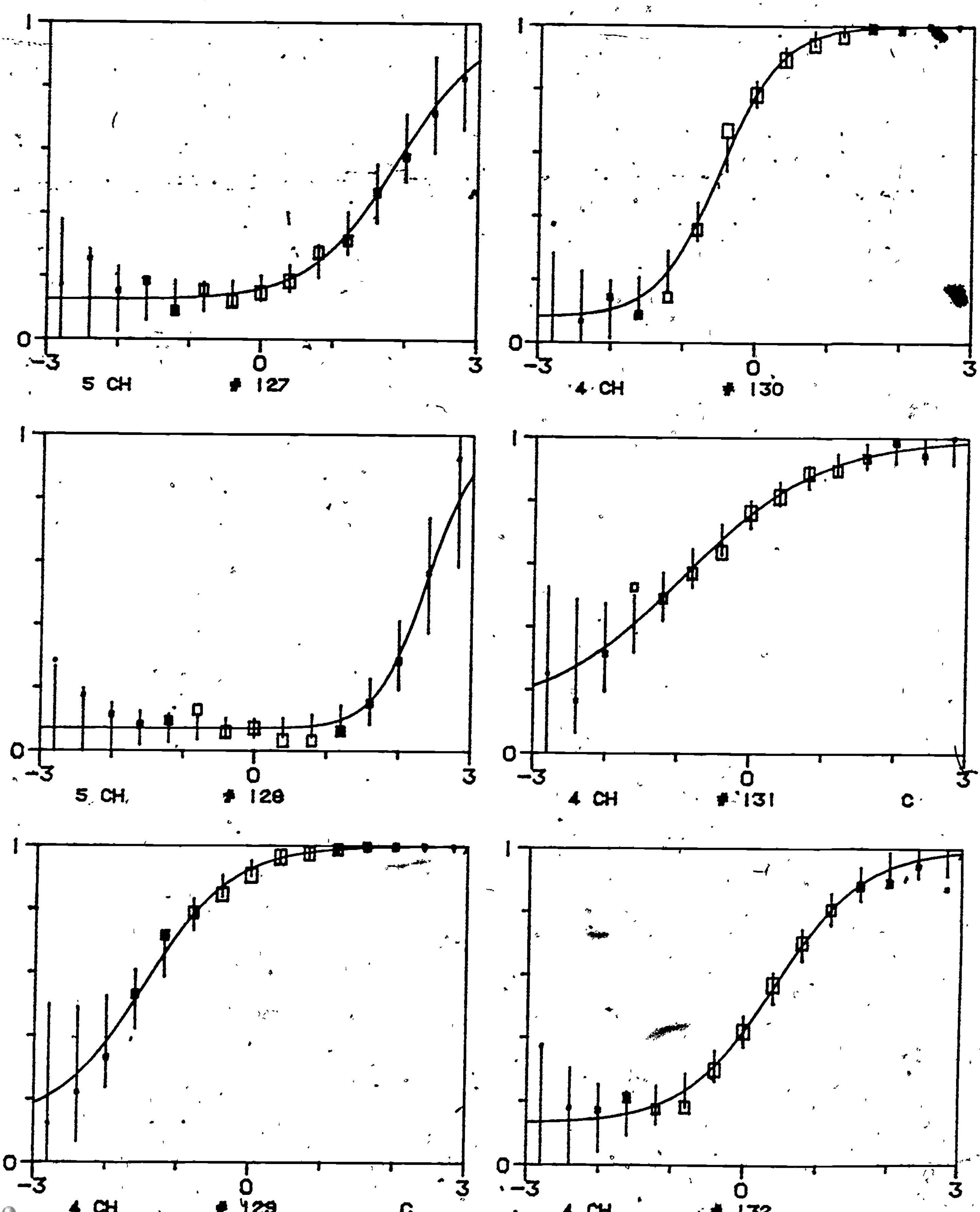


Figure 30: SAT Second Old Form Item Ability Regression Plots

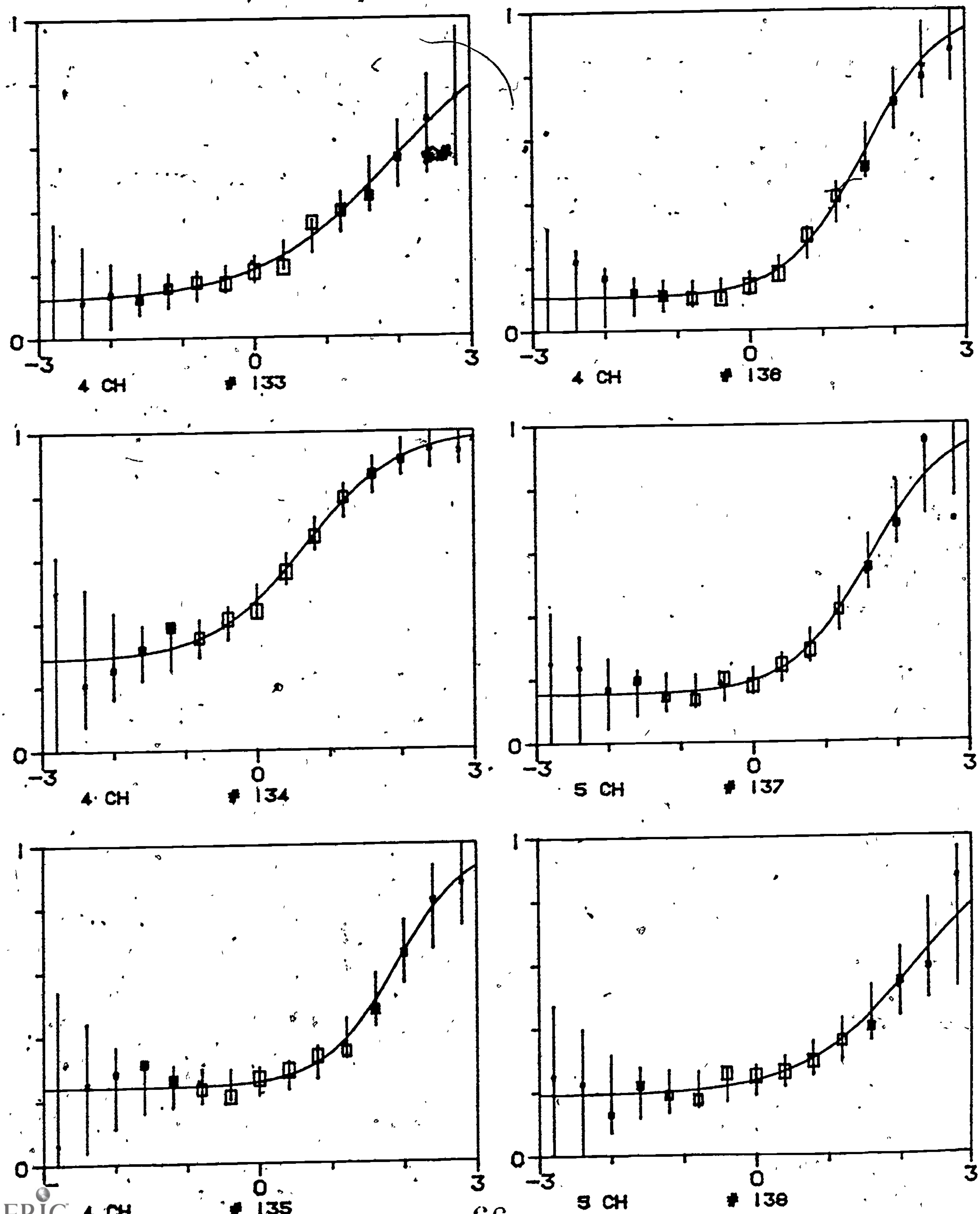
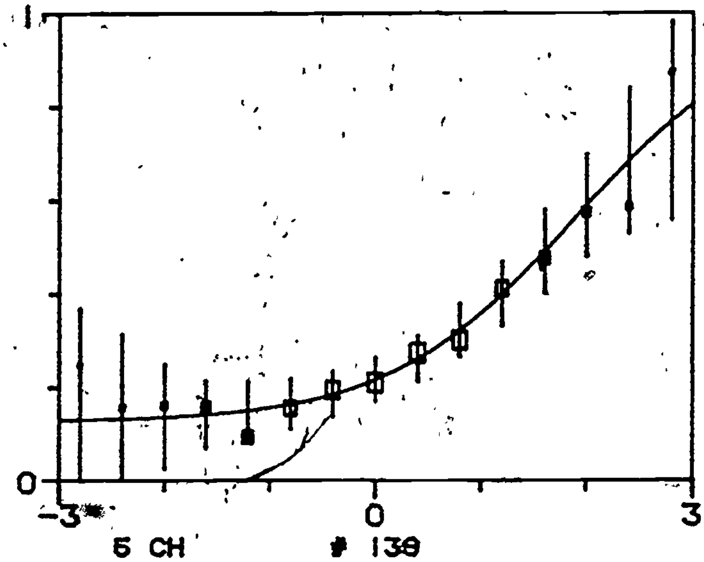
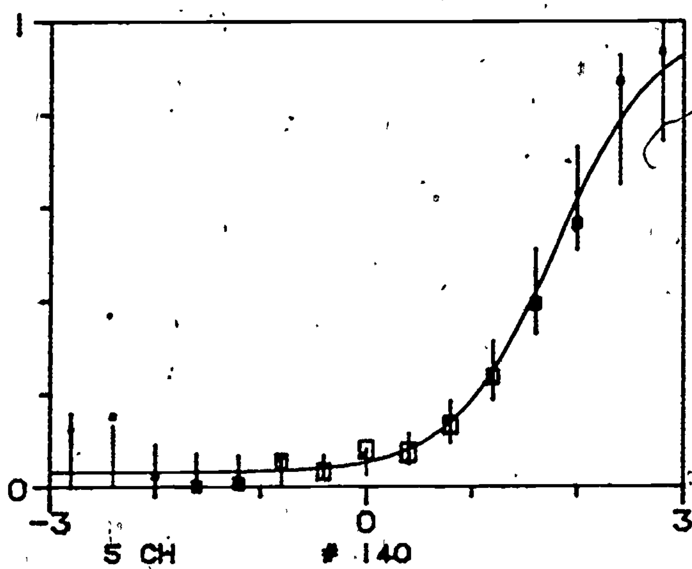
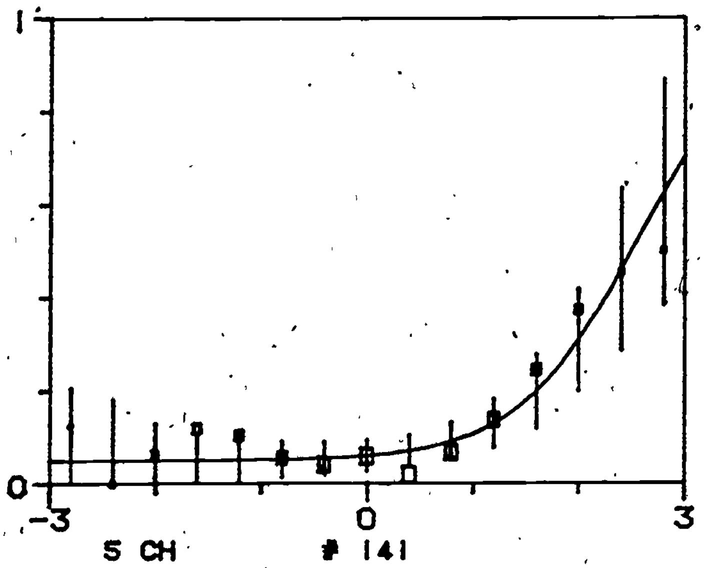


Figure 31: SAT Second Old Form Item Ability Regression Plots



It should be noted, when inspecting the information in Table 7, that the interval contributions have signs attached to them. One of the weaknesses of the use of chi-square like statistics to assess model-data fit is that the statistics do not take into account whether or not the deviations of observed values from expected values are in a single direction. This is particularly important as an erratic pattern of pluses and minuses may be indicative of random fluctuations in the data, whereas a number of pluses (or minuses) ocurring in a row indicate that the empirical regression is consistently above (or below) the estimated regression for specific ranges of the ability continuum.

For each item, the information obtained from the $Q_1'$ statistic was used in conjunction with the item ability regression plot (Figures 6-29) obtained for that item to determine goodness of fit. The process consisted of first, inspecting the item ability regression plot to form an opinion as to how well the item fit the data. If the mid-points of a series of boxes, representing the empirical regression, fell outside the $\pm 2\sqrt{PQ/N_j}$ interval centered on the response function, the value of $Q_1'$ (particularly the contributions of the ability level intervals and the sign of these contributions) was inspected to form a judgement as to whether the item was truly poorly fitting. Secondly, the process was reversed and the $Q_1'$ statistic for a particular item was inspected and used as a flagging mechanism to determine if the item ability regression plot warranted close inspection. If the value of the statistic was large in relationship to values obtained for other items, the contributions of the ability level intervals was studied. An opinion was formed regarding the goodness of fit of the item and an attempt was made to verify this opinion by examination of the item ability regression

plot. Using this back and forth procedure, the authors were able to isolate 24 items for which the goodness of fit to the three parameter model was judged questionable. These items are indicated in Table 7 by an asterisk placed next to the item number. Several of these items will be discussed in the following paragraphs.

Examination of the data contained in Table 7 indicates that the value of $Q_1'$ for item 1 (PSAT/NMSQT Form 1) is 65.96. Although selected ability level intervals ($\theta = 1.6$, $\theta = .8$) are contributing heavily to the value of this statistic, other intervals are also making contributions of considerable size. The fairly consistent pattern of pluses and minuses is indicative of the ranges on the ability continuum for which the item fits poorly. The value of the c parameter for this item was set to .067. Inspection of the item ability regression plot shown in Figure 6 indicates that had the c parameter been allowed to assume a lower value, fit would have been improved for lower ability levels. This would not, however, have improved fit at the upper end of the ability continuum.

The item ability regression plot for item 5 (PSAT/NMSQT Form 1), given in Figure 6, indicates that the empirical regression falls above the estimated regression for ability levels below -2. On the other hand, for ability levels just below $\theta = 0$ and ability levels above $\theta = +2$, the empirical regression falls below the estimated regression. The value of the $Q_1'$ statistic for this item (given in Table 7) is 32.71. An examination of the contributions for the ability level intervals verifies the pattern that is displayed by the item ability regression plot, i.e., the observed data does not fit a monitonically increasing function very well. One might question why item 5 was selected and item 4 (PSAT/NMSQT Form 1) was not, given that

the value of $Q_1'$ is slightly larger for item 4 than for item 5. There are two reasons for this. First, it would appear, from the item ability regression plots, that item 4 fits the model better than item 5 (fewer boxes fell out of the interval defined by $\pm 2\sqrt{PQ/N_j}$). Secondly, from examination of the contributions of the ability level intervals to the overall value of $Q_1'$ for item 4, it is noteworthy that only one category ($\theta = 2.8$) is contributing over half of the value of the statistic.

The item ability regression plot for item 18 (PSAT/NMSQT Form 1) shown in Figure 8 indicates some poor fit at the extremes of the ability continuum, however, from the appearance of the plot it would not be expected that the value of $Q_1'$ would be as large as 223.64 (the largest value obtained for any item in this study). Examination of the individual ability level interval contributions to the overall chi-square value indicates that the interval containing responses of examinees with $\theta$ greater than +3 has a value of -145.15. Of the twelve examinees in this interval, only one answered the item incorrectly inflating the contribution of this interval to the overall statistic unreasonably. The item was judged to be poorly fitting because of the fairly large contributions of the remaining intervals and the consistent pattern of the plus and minus signs for these intervals.

The final example of a poorly fitting item that will be discussed is item 44 (PSAT/NMSQT Form 1). The item ability regression plot for this item given in Figure 13 clearly shows the non-monotonicity of the data. The fact that this is a poorly fitting item may be verified by examination of the chi-square information given in Table 7. No single ability level interval seems to be contributing unreasonably to the overall value of 69.03. Also, the consistent pattern of pluses and minuses reflects the non-monotonicity observed in the item ability regression plot.

In summary, a few general comments can be made. First, the majority of items judged to have questionable fit to the three-parameter model were PSAT/NMSQT items. It is likely that this is an artifact of the particular forms chosen. Were the experiment to be repeated with four different forms, the greatest number of poorly fitting items might have been found in the SAT forms.

Secondly, it appears as though the majority of PSAT/NMSQT and SAT items judged as poorly fitting basically exhibited poor fit for the extremes of the ability continuum, particularly for lower ability levels. This is most probably related to the problem of obtaining an adequate estimate of the psuedo-guessing parameter of the three-parameter logistic model.

## Discussion

The equating models examined in this study are based on a number of assumptions, most of which are violated to some extent. One of the assumptions underlying all of the equating models is that the two tests to be equated are unidimensional (Morris, in press). Because the IRT equating model assumes unidimensionality on the item level whereas the linear and equipercentile models used for this study only assume unidimensionality of test scores, one might expect violations of this assumption to have a more serious effect on the IRT equating results. However, unidimensionality is a necessary condition for the establishment of a single common metric regardless of the equating model. Thus, it is difficult to say what the implications of violation of this assumption are for any of the equatings.

The equatings are certainly affected in a differential manner by availability of data. Insufficient data at low ability levels generally leads to problems in estimating the psuedo-guessing parameter for the three-parameter logistic model. It is also difficult to estimate discrimination parameters and difficulty parameters for very easy or very difficult items if adequate data does not exist. Equipercentile equating methods are particularly sensitive to lack of data at the extremes of the ability continuum. Scarcity of data at these extremes can lead to serious problems in establishing an equitable relationship for high and low scores on the two test forms. The linear equating method, which is based on estimated means and standard deviations, can also be affected by the influence of outlying values on these estimates. This was probably not a problem for the present study, given that sample sizes were quite large and should have been sufficient to produce stable estimates.

Differences in test reliability must also be considered. Lord (1980, Chapter 13) states that in order to accurately equate two tests, i.e. produce scores on two tests such that it is a matter of indifference to examinees which test they take, the tests must be strictly parallel and perfectly reliable. It is difficult to predict how the various equating methods are affected by differences in test reliability (certainly a problem for this study, given the differences in test length). However, the methods based on true-score estimates (IRT and Levine Unequally Reliable) should be least affected by the problem.

The lack of parallelism between the tests to be equated has particular implications for the linear method which requires, in order to adequately describe the relationship between scores on two forms of a test, that the distributions of these scores differ only in their means and standard

deviations. Lack of parallelism between two tests generally results in a curvilinear relationship between raw scores necessitating an equating method such as IRT or equipercentile to produce accurate results.

One must also consider the problems posed by the differences in the level of ability of the PSAT/NMSQT and SAT groups. The purpose of anchor test designs is to provide a mechanism to adjust for these differences. Marco, Petersen and Stewart (1979, in press) report that as group differences become more pronounced, the quality of the equating suffers. The results of their study indicate that IRT methods, based on item parameters that are group invariant, may produce better results in this situation.

In the absence of an adequate criterion upon which to judge the equatings, one must rely heavily upon the arguments described above to draw any conclusions regarding the results of the present study. It is apparent, from examination of the equating results, that the relationship between the raw scores on the PSAT/NMSQT and SAT is curvilinear, particularly for the upper and lower ends and somewhat for the middle of the score range. Therefore, the Tucker and Levine Unequally Reliable linear methods are probably inadequate for describing the raw score relationship between the PSAT/NMSQT and the SAT. Because the equipercentile method is so sensitive to scarcity of data at the extremes of the score range, the IRT method would appear to describe the curvilinear relationship in a more appropriate manner. This is most probably true for the upper end of the score range. A decision as to which equating method (IRT or equipercentile) is most effective at the lower end of the score range is difficult to make. Both methods are affected by scarcity of data; accuracy of the IRT method depends on how adequately the c parameters were estimated. As indicated by the goodness of fit assessment,

problems with the estimation of this parameter contributed, in some cases, to the questionable fit of items.

It is probably safe to conclude from the goodness of fit assessment that, in spite of some problems related to the estimation of the psuedo-guessing parameter, the data fit the three-parameter model fairly well. Certainly, this aspect of the study suffered from some methodological problems. The manner in which the ability continuum was divided into intervals for the chi-square statistic was a particular problem. Some method of accounting for within interval variance, such as subtracting the variance of the predicted proportion passing the item for a specific interval from the denominator of the chi-square statistic (Wright and Mead, 1977), should probably have been considered. Ultimately, the decision as to whether or not an item fit the three-parameter model was a judgmental one. This could not be avoided, given the present state of the art.

Several of the problems encountered in the study are currently being investigated at Educational Testing Service. The LOGIST program has recently been revised to improve the estimates obtained for the psuedo-guessing parameter. Although it was not possible to repeat the entire study (specifically, the equatings), item parameter estimates and item ability regression plots were obtained for the same set of data using the revised program and compared to those used for the present study. Substantial improvement was found in the estimates obtained for the c parameters for selected items.

It is planned to revise the goodness of fit statistic in the near future so that some of the problems encountered with its application in this study will be avoided. Hopefully systematic use of this statistic in conjunction with item ability regression plots will provide information regarding model-data fit that can be studied on a continuing basis.

To summarize, the conclusions that can be drawn from the present study regarding the appropriateness of the three-parameter model for a vertical equating situation and the goodness of fit of the data to this model are limited. There are two reasons for the limitations: (1) the lack of a criterion for judging the equatings; and (2) the subjective nature of the goodness of fit assessment. Further research is certainly needed. An important direction for this research to take is the direct assessment of the affect of the number and degree of poorly fitting items on IRT equating results. This assessment should take into account such variables as equating design, differences in group ability and lack of parallelism between the tests to be equated.

More generally, the extent to which IRT methods are being implemented in practical testing situations has dramatically increased over the past several years. It is important that researchers interested in the quality of the results of the many applications rigorously pursue the question of goodness of fit and the possible implications of lack of fit for the specific applications they are interested in.

REFERENCES

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Cook, L. L., Dunbar, S. A., and Eignor, D. R. IRT equating: A flexible alternative to conventional methods for solving practical testing problems. Paper presented at the annual meeting of AERA, Los Angeles, 1981.

Divgi, D. R. Does the Rasch model really work? Not if you look closely. Paper presented at the annual meeting of NCME, Los Angeles, 1981.

Hambleton, R. Latent Ability Scales: interpretation and uses. In S. Mayo (Ed.), New Directions for Testing and Measurement; Interpreting Test Performance, No. 6. San Francisco: Jossey-Bass, 1980.

Kolen, M. J. Comparisons of traditional and item response theory methods for equating tests. Journal of Educational Measurement. 1981, 18, 1-11.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Marco, G. L., Petersen, N. S. & Stewart, E. E. A test of the adequacy of curvilinear score equating models. Paper presented at the 1979 Computerized Adaptive Testing Conference, Minneapolis, 1979.

McKinley, R. L., and Reckase, M. D. A comparison of the ANCILLES and LOGIST parameter estimation procedures for the three-parameter logistic model using goodness of fit as a criterion. Research Report 80-2. Arlington, VA: Personnel and Training Research Programs, ONR, 1980.

Morris, C. On the foundations of test equating. In P. Holland (Ed.) Proceedings of the ETS Research Statistics Conference on Test Equating. New York: Academic Press, in press.

Panchapakesan, N. The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago, 1969.

Petersen, N. S., Cook, L. L., and Stocking, M. L. Scale drift: A comparative study of IRT versus linear equating methods. Paper presented at the annual meeting of AERA, Los Angeles, 1981.

Petersen, N., Marco, G., & Stewart, E. E. A test of the adequacy of linear score equating models. In P. Holland (Ed.), Proceedings of the ETS Research Statistics Conference on Test Equating. New York: Academic Press, in press.

Rentz, R. R., and Rentz, C. C. Does the Rasch model really work? A discussion for practitioners. ERIC Report No. 67. Princeton, NJ: Educational Testing Service, 1978.

Rentz, R. R., and Ridenour, S. E. The fit of the Rasch model to achievement tests. Paper presented at the annual meeting of EERA, Williamsburg, VA: 1978.

Slinde, J. A. and Linn, R. L. Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 1977, 14, 23-32.

Slinde, J. A. and Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Slinde, J. A. and Linn, R. L. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, 1979, 16, 159-165.

Stocking, M. Personal communications, 1980.

Wood, R. L., and Lord, F. M. A user's guide to LOGIST. Research Memorandum 76-4. Princeton, NJ: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., and Lord, F. M. LOGIST - A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D., and Stone, M. H. Best test design. Chicago, IL: Mesa Press, 1979.